Functional Exploration of Antisense Long Non-Coding RNAs Containing Transposable Elements: A Bioinformatics Approach

Thesis

Functional exploration of antisense long non-coding RNAs containing transposable elements: a bioinformatics approach

A thesis submitted to the Open University of London for the degree of

Doctor of Philosophy

by

Veerendra P Gadekar

Master of Science in Bioinformatics, Manipal University,

Manipal- Karnataka, India

Stazione Zoologica Anton Dhorn, Naples, Italy

The Open University, London, United Kingdom

Director of studies: Dr. Remo Sanges, Ph.D.

External Supervisor: Dr. Stefano Gustincich, Ph.D.

Co-supervisor: Dr. Paolo Sordino, Ph.D.

January, 2016

DATE OF SUBMISSION : 29 JANUARY 2016

DATE OF AWARD : 30 SEPTEMBER 2016

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Long non-coding RNA (lncRNAs) show a wide range of regulatory functions at the transcriptional and post-transcriptional levels both in the nucleus and cytoplasm. Recently, antisense lncRNAs (ASlncRNAs) were reported to up-regulate protein synthesis post-transcriptionally through a mechanism depending on an embedded inverted SINE B2 and 5' overlap to the target mRNAs. Such ASlncRNAs are also referred as SINEUPs. Synthetic SINEUPs with identical modular organization were also demonstrated to exert the same activity suggesting a functional relationship between SINE repetitive elements and ASlncRNAs. In order to gain a broader insight on the contribution of transposable elements (TEs) in the sequence composition of ASlncRNAs, I have developed a bioinformatic pipeline that can identify and characterize transcripts containing TEs and analyze TEs coverage for different classes of coding/non-coding sense/antisense (S/AS) pairs. I aimed at identifying if the functional activity of SINEUPs could be a widespread phenomenon across multiple similar natural ASlnRNAs in the transcriptomes of the extensively studied model organisms that have a well annotated catalog of lncNRAs. From my initial analysis I identified human and mouse are the two species that showed a significant coverage enrichment of SINE repeats among ASlncRNAs. I further performed several functional enrichment analysis for the sense coding genes overlapping to ASlncRNAs taking into consideration of different characteristics of the 5' binding domain and the 3' embedded SINE repetitive elements. This permitted me to identify the effect of these modular features over the functional associations of sense coding genes. The results of the analysis showed that the products of coding genes associated to ASlncRNAs containing SINEs are significantly enriched for mitochondrial localization. Further, to determine if these ASlncRNAs could exert SINEUP-like activity during stress, I analyzed the

data from a published custom microarray experiment study, that were associated to the polysome fractions of MRC5 cell lysates in control and oxidative stress condition. The results revealed that the ASlncRNA carrying inverted or direct SINE repeats and their corresponding sense coding genes do not show any significant differential polysome loading in stress with respect to normal conditions, which is not a desired characteristic of a potential SINEUP. However, ASlncRNAs with inverted and direct SINE repeats corresponding to high translating polysome fractions showed a significantly higher ratio of means for RNA levels in stress over control, in contrast to noASlncRNA. This suggests that the ASlncRNA containing SINE elements are the key RNA molecules that are active during stress, although to determine if they are also involved in the increased polysome loading of their respective sense coding mRNAs, there is a need of further experimentation and exploration. Altogether, the work presented in this thesis provides a novel bioinformatics approach to study transcriptome-wide ASlncRNAs containing TEs and their functional association over the sense coding genes, and discover new significant functional features of ASlncRNA to be biologically validated.

Acknowledgments

Firstly, I would like to thank the Open University (OU) and Stazione Zoologica Anton Dohrn (SZN), Naples, for giving me the opportunity and funding to pursue my Ph.D. I would like to express my sincere gratitude to my director of studies Dr. Remo Sanges for the immeasurable amount of support and guidance he has provided throughout my Ph.D. His motivation and supervisory role has helped me to grow as a research scientist. Thank you Remo! for showing enormous patience towards me during this whole time of research and writing of this thesis. Your words of encouragement and advice on my research and career have been priceless.

I would also like to take this opportunity to thank my viva examiners - Dr. Paul Kersey (EBI, UK), Dr. Brunella Franco (TIGEM, Italy) and Dr. Marizio Ribera D'alcalà (SZN, Italy) for their very helpful comments and suggestions. My sincere thanks also goes to Dr. Raffaella Casotti, the chair of the board of examiners, for organizing my Ph.D. viva. I also thank her for being very supportive and helpful throughout my Ph.D., as the coordinator of the Ph.D. programs at SZN. I would also like to acknowledge the support and invaluable assistance of Dr. Gabriella Grossi, the secretary of OU Ph.D. programs at SZN, in resolving any kind of issues and documentations related to my Ph.D.

I am very thankful to my external supervisor Dr. Stefano Gustincich (SISSA, Italy), for his timely visits and valuable advice throughout my research. I am extremely grateful to receive his comments and questions that have helped me to improve my presentation skills. I also thank my co-supervisor Dr. Paolo Sordino (SZN) for being kind and supportive during all the time of research and sharing his insightful suggestions.

My heartfelt thanks also goes to Mr. Prashantha Hebbar (DDI, Kuwait), he was my tutor during my master's degree. His motivations, advice and support on many occasion were invaluable and inspired me to pursue a career in bioinformatics and research.

A good support system is important to survive the ups and downs of three years of Ph.D. I was lucky to have my good friends and lab-mates for this. I thank Francesco, Giuseppe, Swaraj and Massimiliano for being there and cheering me up whenever I needed. I will never forget the innumerable coffee breaks we have taken together when we discussed science and philosophy. I would also like to thank all my batch-mates at SZN, with whom I share a lot of common experience of growing up as a researcher.

I can't thank enough to my dear friend Harpreet, who has been my weekend gaming partner! I just cannot forget hours of playing PES and COD in his Xbox. Being in such a special city like Naples became even more memorable to me because of all my cheerful friends. I would like to thank Pacos, Veer, Christopher, Mauro, Gianni for accompanying me to the sightseeing, in and around Napoli and for BBQ Hangouts!

My special thanks goes to a very sweet Italian family with whom I stayed during this time. I am thankful to Maria, Paolo and Roberto for treating me as a family member and teaching me to cook Italian cuisines! I would always cherish the time we spent together.

Finally, I would like to thank my family. Words cannot express how grateful I am to my parents for their love, support and all of the sacrifices that you've made for me. I am very

grateful to my elder brother who is also my role model and my two lovely elder sisters for being so supportive during my away time from home. The last word of acknowledgment I have saved for my dear wife Sweta. Thank you for having unwavering love and support towards me. I really appreciate the extra-special care you have shown during this whole time of my Ph.D., it helped me to stay motivated until the writing of this thesis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIR	Antisense Igf2r RNA
ASlncRNA	Antisense lncRNA
BACE1	Beta-secretase-1 gene
BP	Biological process
CAGE	Cap analysis gene expression
CC	Cellular component
Cdk6	Cyclin-dependent kinase 6
CDS	Coding Sequences
ChIP-seq	Chromatin immunoprecipitation assays with sequencing
CHO-K1	Chinese hamster ovary cells
CNVs	Copy number variations
COLDAIR	Cold Assisted Intronic Non-coding RNA
CRISPR-Cas9	Clustered, Regularly Interspaced, Short Palindromic Repeat associated protein 9 endonuclease
CTCF	CCCTC-binding factor
CTNNB1	Catenin Cadherin-Associated Protein, Beta 1
dATG	Downstream ATG codon
DHFR	Dihydrofolate reductase
DHS	DNaseI hypersensitive sites
DMSO	Dimethyl Sulfoxide
Dnase-seq	DNase I hypersensitive sites sequencing
E2F1	E2F Transcription Factor 1
E2F4	E2F Transcription Factor 4
eIF3	Eukaryotic initiation factor 3
ERVs	Endogenous Retrovirus
ESTs	Expressed Sequence Tags

FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements
gadd7	Growth-arrested DNA damage-inducible gene 7
GIS	Gene Identification Signatures
GO	Gene Ontology
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4me3	Histone 3 lysine 4 trimethylation
HEK	Human Embryonic Kidney cells
hnRNAs	Heterogeneous nuclear RNAs
HOTAIR	HOX transcript antisense RNA
HOTTIP	HOXA transcript at the distal tip
HOXC	Homeobox gene cluster C
HOXD	Homeobox gene cluster D
HuR	Human antigen R proteins
Igf2r	Insulin-Like Growth Factor 2 Receptor
JUNB	Jun B proto-oncogene
LCRs	Locus control regions
lincRNAs	Long intergenic non-coding RNA
LINEs	Long Interspersed Elements
lncRNAs	Long non-coding RNAs
LTR	Long Terminal Repeats
MAML1	Mastermind-Like 1
MEF2C	Myocyte Enhancer Factor 2C
MF	Molecular function
miRNAs	Micro-RNAs
MIRs	Mammalian-wide interspersed repeat
MN9D	Fusion of embryonic ventral mesencephalic and neuroblastoma cells
Mnase-seq	Micrococcal nuclease
MRC5	Human fetal lung fibroblast cells
mTORC1	Mechanistic target of rapamycin, complex1

mTP	Mitochondrial targeting signals peptide
ncRNA	Non-coding RNA
NEAT1	Nuclear Enriched Abundant Transcript 1
noASlncRNA	LncRNAs without antisense overlap
p70S6K1	p70 ribosomal S6 kinase 1
PETs	Paired-end ditags
piRNAs	Piwi-interacting RNAs
PRC2	Polycomb repressive complex 2
RACE	Rapid amplification of 5′ cDNA ends
rCNEs	Regionally conserved non-coding elements
RGNs	RNA-guided nucleases
RNAi	RNA interference
RRBS	Reduced representation bisulphite sequencing
S/AS	Sense-antisense transcript pair
SER3	3-phosphoglycerate dehydrogenase gene
shRNAs	Sharp hairpin-shaped RNAs
SINEs	Short Interspersed Elements
siRNAs	Small-interfering RNAs
SMD	Staufen1 (STAU1)-mediated messenger RNA decay
SmRNA ATG	Sense mRNA with ATG overlap
SmRNA noATG	Sense mRNA without ATG overlap
sncRNA	Small non-coding RNA
snoRNAs	Small nucleolar RNAs
SNPs	Single nucleotide polymorphism
snRNAs	Small nuclear RNAs
SP	Secretory pathways
SRG1	SER3 regulatory gene 1
SRP	Signal recognition particle
SR	SRP receptor

TDP-43	TAR DNA-binding protein
TEs	Transposable Elements
TFBS	Transcription Factor Binding Sites
TIS	Translation Initiation Site (or ATG codon)
TSS	Transcription Start Site
TOP	Terminal Oligopyrimidine Tract
Uchl1	Ubiquitin Carboxyl-Terminal Esterase L1
UCRs	Ultra-conserved Regions
UPF1	Up-frameshift protein 1
UTR	Untranslated region
Uxt	Ubiquitously expressed transcript gene
Xic	X inactivation center
Xist	X-inactive specific transcript
4EBP1	eIF4E-binding protein 1

Chapter 1

Introduction

1.1. Long non-coding RNAs in the post-genomic era

1.1.1. Brief history of the genomic era

The human genome project (HGP) is one of the greatest endeavors in the field of genomics and molecular biology that led to the dawning of genomic era. It was launched as a large scale international collaborative research program with a goal to sequence the complete human genome, aiming for understanding how the genetic information determines the development, structure and function of the human body. The HGP led by the international human genome sequencing consortium (IHGSC) also sought to develop tools to obtain and analyze the sequence data and to make this information widely available. This would also lead to the advancements in understanding of how variations within our DNA sequence could cause disease and how such diseases could be cured or prevented using "personalized" medicine. The accomplishment of all these goals would certainly mean a big leap of humankind towards the comprehension of molecular nature of life. Today, by achieving the initial goal of sequencing the human genome and making it publicly available through the three primary portals: the University of California Santa Cruz (UCSC), Ensembl (of the European Bioinformatics Institute; EBI) and the NCBI (National Center for Biotechnology Information; part of National Institutes of Health), the HGP has already set a platform for new studies which are presently deciphering the evolution of eukaryotic genomes and factors which are involved in reshaping and regulating their genomic activities (International Human Genome

Sequencing Consortium, 2001). However, the IHGSC were not alone to achieve this important milestone but were accompanied by an independent group represented by Celera Genomics led by Craig Venter (Venter et al., 2001). The genome assemblies published by both contained similar amount of genomic sequences and gaps that were filled in later releases. For sequencing, IHGSC followed the *hierarchical shotgun* approach, whereas Celera Genomics used the *whole genome shotgun* sequencing. Several comparative analysis for the two genome sequence assemblies by IHGSC and Celera Genomics have been published revealing that both the assemblies consisted of approximately equal number of predicted genes. However, they showed very little overlap for novel predicted gene sets (Hogenesch et al., 2001). Taken together, the IHGSC's approach of *hierarchical shotgun* sequencing gained more importance in terms of better state of assembly, whereas the *whole genome shotgun* sequencing was considered as a challenging strategy for sequence assembly (Li et al., 2003).

After the successful achievement of the initial goal of sequencing the complete human genome, next step to extract hidden information within ~3 billion nucleotides constituting the layout for functional RNA and protein molecules within a cell would be a much bigger challenge. This was discerned by the HGP well before the outset of human genome sequencing, and an approach of comparative genomics was thought to be useful in discovering the hidden information from the sequence data. As a result, along with the sequencing of human genome well underway by 1999, a concerted effort to sequence the entire mouse genome was organized by the Mouse Genome Sequencing Consortium (MGSC) (Mouse Genome Sequencing Consortium, 2002). The initial comparative analysis of human and mouse genome revealed that the mouse genome is 14% smaller than human genome. However, 90% of both the genomes were identified to be maintaining conserved syntenic regions and were

estimated to contain about 30,000 protein-coding genes (Mouse Genome Sequencing Consortium, 2002). As per Ensembl release 82 *(September 2015)* the total number of protein-coding genes in human (hg38) and mouse (mm10) genomes are 22,017 and 22,158 respectively.

Prior to the sequencing of human and mouse genomes, the HGP also initiated sequencing of simpler eukaryotes that are used in laboratories as model organisms. These included *Saccharomyces cerevisiae (yeast)* with a genome size of 12 megabases containing ~6000 coding genes (Goffeau et al., 1996), *Caenorhabditis elegans (worm)* with 97 megabase genomic sequence containing ~19,000 coding genes (The C. elegans Sequencing Consortium, 1998), *Drosophila (fly)* with a genome size of ~120 megabases predicted to carry ~14,000 coding genes. *Drosophila* genome sequencing was initially led by *Berkley Drosophila Genome Project* and the *European Drosophila Genome Project.* Later with the collaboration of Celera Genomics it became the first genome to be sequenced using the *whole genome shotgun* sequencing approach (Adams et al., 2000; Ashburner & Bergman, 2005). In parallel with the HGP's large scale efforts, the first genome of plant kingdom represented by *Arabidopsis thaliana* was also sequenced by the *Arabidopsis genome initiative in 2000. A.thaliana* was considered an important model system because its genome sequence could reveal the genetic differences between plants and other eukaryotes. *A.thaliana* has a genome size of ~125 megabases predicted to carry 25,498 coding genes (The Arabidopsis Genome Initiative, 2000).

Undoubtedly, the large-scale genome sequencing efforts and the comparative genomic analysis of first six sequenced eukaryotic genomes established a platform for further discoveries. The initial observations were already astonishing, the number of protein-coding genes in

sophisticated organism such as human, with more than 200 distinct cell types were almost equal to mouse, and only slightly higher than simpler eukaryotes such as worm and fly that contain as few as 28 and 64 distinct cell types respectively (Schad, Tompa, & Hegyi, 2011; Liu, Mattick, & Taft, 2013). Considering the number of distinct cell types as a proxy for organismal complexity (Chen et al., 2014) and their apparent lack of correlation between the number of protein-coding genes in different organisms suggests, that the number of protein-coding genes might not be the only factor which determines organism's complexity. Instead, there should be other important aspects of eukaryotic genomes which are involved in several crucial roles that correlate well with distinct cell types to governs various cellular processes among diverse eukaryotes. The identification of the factors involved in the determination of organismal complexity has been a hot topic of research from a long time.

1.1.2. Paradoxes associated with organismal complexity

1.1.2.1. C-Value paradox

One might expect complex organisms to have larger genome size to correlate with their distinct cell types and sophisticated morphology. This would imply that the complex organisms should contain more DNA per cell as they would require more functional genes to correlate well with their apparent complexity. However, even the primeval studies showed that this is not the case, in fact many apparently simpler organisms could have over a thousand times more DNA than complex multicellular organisms or multiple organisms with similar complexity level could widely differ in their DNA content (Hilder et al., 1981; Mirsky & Ris, 1951). This disjunction in the DNA content between the simple and complex organisms was referred as the **C-value paradox** (Thomas. C. A, 1971), where the C-value stands for the

amount of DNA per haploid set of chromosomes, usually measured in millions of base pairs (Mb) or picograms (pg) (Swift, 1950).

Many hypotheses have been proposed in literature for the explanation of C-value paradox suggesting, the bulk of DNA has adaptive significance independent of its protein-coding function. One of such explanations included the introduction of nucleoskeletal DNA (S-DNA) concept, according to which the major portion of the DNA is composed of nucleoskeletal DNA that does not encode for proteins (Cavalier-Smith, 1978) but exists to render its nucleoskeletal role in determining the nuclear volume in the cell and might affect features such as the rate of cell division and development. This implies that the changes in genome size may be adaptive. On the other hand, studies claimed that the accumulation of DNA is largely non-adaptive, instead they are in selective pressure and only a small portion of the eukaryotic genome sequence is conserved (Marcus, 2005).

The most widely accepted explanation of the C-value paradox concerned a different line of thinking where the genomes carry a fraction of DNA that does not encode for proteins hence are biologically trivial in the development of organism, with very little or no adaptive advantage for the organism. They were also addressed as the non-coding DNA or "junk" DNA (Ohno, 1972). Some genomes carry the non-coding fraction of DNA more than others, and some genomes carry quite a lot of it. The labeling of the non-coding fraction of DNA as "junk" DNA by Ohno seemed to settle down the C-value paradox for quite some time (Eddy, 2012).

1.1.2.2. N or G-value paradox

However, the comparison of the predicted number of coding genes among first six completely sequenced eukaryotes once more arose the question, "what determines the organismal complexity?", as it was observed that the simplest eukaryote such as worm that possess 28 distinct cell types carry at maximum only one-third number of genes less than human. Some chose to call this as **N-value paradox**, where neither DNA content nor gene number could be used to specifically address the organismal complexity. Therefore, researchers started to believe that the number of transcripts a genome could express would probably be a more effective measure to be associated with the organismal complexity (Jean-Michel Claverie, 2001). This means that the organismal complexity should be independent of the number of coding genes or the measurement of DNA content. Instead, should depend upon the transcriptional outputs and multiple other properties of higher eukaryotic transcriptomes (Harrison et al., 2002) revealed in later studies. For example, the alternative splicing of mRNAs, alternative poly-adenylation, complex promoters (Gagniuc & Ionescu-Tirgoviste, 2012), and gene regulatory networks.

Others called the lack of correspondence between the gene number and organismal complexity as **G-value paradox** (Hahn & Wray, 2002) and started to study alternative aspects of eukaryotic genomes such as cis-regulation, multi-functional proteins, post-translational modifications (Alberts et al., 2002), and gene duplication (Friedman & Hughes, 2001). Gene duplication itself was identified as a major evolutionary force that has acted upon *C. elegans* genome, resulting into the formation of about one-third (32%, >6,100 genes) of its total genes through duplication, where the duplicated blocks of genes were intra chromosomal and a single duplicated block were found to contain ~21 genes (Friedman & Hughes, 2001; The C.

elegans Sequencing Consortium, 1998). Several theories have been proposed to explain gene duplication, for example, Ohno (1970) theorized the gene duplication as a scenario where mutation replicates a single gene into two copies, where one gene duplicate will experience relaxed selection and will accumulate mutations. And the other gene duplicate will undergo purifying selection for the ancestral function by avoiding the accumulation of deleterious mutations. By this mechanism, the evolutionary fate of most gene duplicates is thought to be degeneration and "nonfunctionalization" through pseudogene formation (Ohno, 1970).. Alternatively, to explain the consequential functional redundancy in the duplicated genes, a theory of "subfunctionalization" has been proposed, where two genes may overlap in some of their functions, but each has at least one unique function (Lynch & Force, 2000). Hence, gene duplication is one of the possible explanations for the unexpectedly large number of genes accounted in *C. elegans* genome, in comparison to *Drosophila* and human (Hodgkin, 2001). Altogether, it is now clear that the eukaryotic genomes are more complex than expected and their understanding would rely on the detailed exploration of various characteristics of eukaryotic genomes.

1.1.2.3. Role of non-coding DNA in organismal complexity

In 2004, Taft and Mattick confirmed that the amount of non-coding DNA per genome is a valid measure of the complexity of an organism (Taft & Mattick, 2004). They analyzed the ratio of the non-coding DNA to the total genomic DNA (ncDNA/tgDNA) for 85 sequenced genomes including prokaryotes and eukaryotes and found a positive correlation in the organismal complexity with the increasing ncDNA/tgDNA ratio. Among all the computed ncDNA/tgDNA ratios, humans held the highest value *(Figure 1.1)*, hence may reasonably be

considered as the most complex organism in the biosphere with large number of distinct cell types, sophisticated brain and body plan.

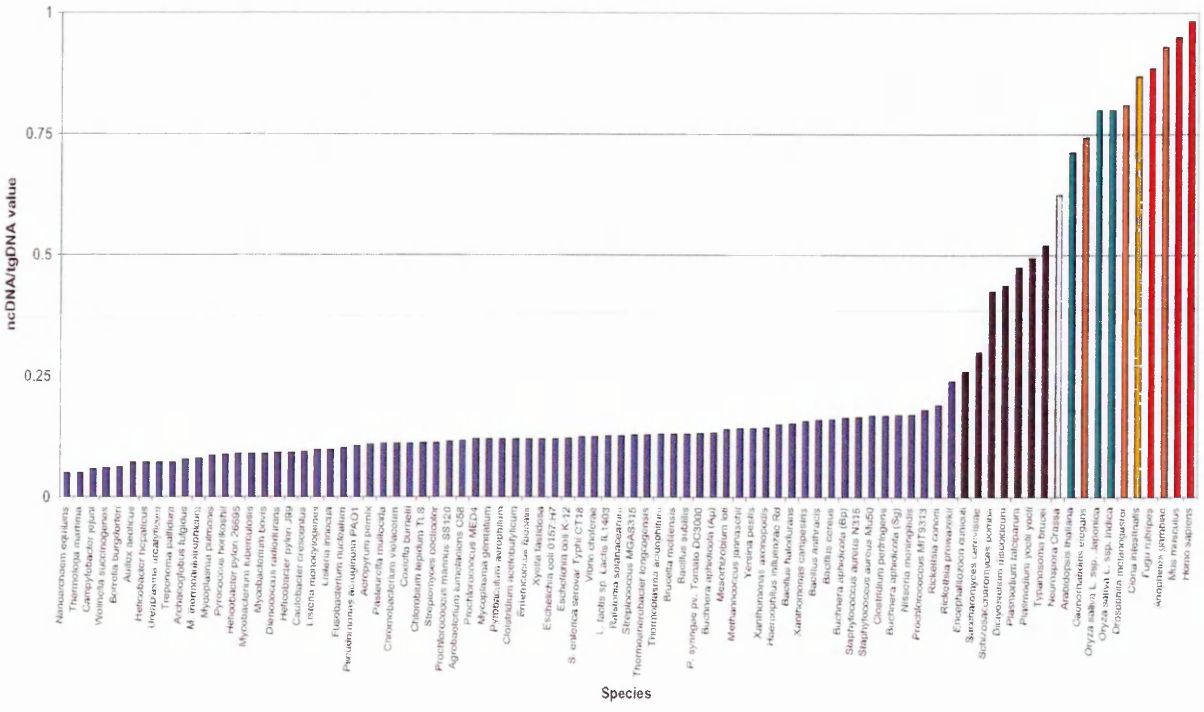


Figure 1.1 | The increase in the ratio of noncoding DNA to total genomic DNA (ncDNA/tgDNA) is shown to correlate with increasing biological complexity. For ease in understanding of phylogeny status of the included organisms, prokaryotes are labeled in blue, unicellular eukaryotes in black, the multicellular fungus Neurospora crassa in gray, plants in green, non-chordate invertebrates in brown, the urochordate Ciona intestinalis in orange, and vertebrates in red. *(Above set of figures are taken from Taft & Mattick, 2004).*

The study of *Taft and Mattick* also suggests that the previously regarded "junk" DNA could in fact be the key player in evolution and development of complex organisms. There are several

lines of study which support this notion. For example, a total of 481 segments of sequence longer than 200 bp length are found absolutely conserved with 100% identity between orthologous regions of the human, rat and mouse genomes, hence are also referred as the ultra-conserved regions (UCRs) (Bejerano et al., 2004) or ultraconserved non-coding elements (UCNEs) (Dimitrieva & Bucher, x`2012). These UCRs are also found to be in a strong positional correlation (synteny) with the transcription factor encoding genes and the genes that encode for the key regulators of development (Sandelin et al., 2004). Hence, the UCRs are suggested to be the good candidates for regulatory elements important in the early stages of vertebrate development. For example, an UCR is reported to be present immediately upstream to the *HoxA7*, and is also described to act as the enhancer of *HoxA7* in human and mouse, but at the same time it is found absent in the zebrafish *Hox* clusters. This interestingly correlates with the fact that the *HoxA7* gene was lost during zebrafish genome evolution (Santini et al., 2003; Spitz, Gonzalez, & Duboule, 2003). These findings suggests for the involvement of non-coding genomic regions in the UCR-mediated molecular events (Harmston, Baresic, & Lenhard, 2013). Similarly, highly conserved non-coding sequences are also identified between human and the pufferfish *Fugu rubripes*, specifically around the set of genes which are related to the developmental processes and transcriptional factors (Woolfe et al., 2005). Regionally conserved non-coding elements (rCNEs) are also identified across multiple vertebrate genomes including mammals and fish. These rCNEs undergo shuffling, where the majority are likely to act as enhancers for the genes involved in either developmental process or regulation of transcription (Sanges et al., 2006). Given that these rCNEs are also found conserved among the highly diverged species such as human and fish, they must therefore be functionally important for vertebrates in the developmental process. Away from the eukaryotic animals, plants such as maize, rice and other diverged members of monocots are also identified as

containing conserved non-coding sequences particularly enriched among the genes with upstream regulatory roles (Inada et al., 2003). Furthermore, a few of the early studies have shown the evidence for the transcription of non-coding sequences in the mouse genome that give rise to functionally active non-coding RNA molecules (H19, Xist, AIR) involved in imprinting and other cellular processes (Bartolomei, Zemel, & Tilghman, Hernals1991; Brockdorff et al., 1992; Lyle et al., 2000). Altogether, these studies suggest that the non-protein-coding sequences in the genomes of eukaryotic organism contain a large amount of regulatory information, indicating their probable role as the key regulators of genomic activities. And further exploration of the transcriptome profile of the organisms should be the first step towards better understanding of the functional involvement of non-coding sequences, because the transcriptome analysis experiments can characterize transcriptional activities of all coding and non-coding transcripts and provide opportunity to perform several comparative analysis for the dissection of the functional roles of lncRNAs.

1.1.3. Major projects venturing transcriptomes

1.1.3.1. FANTOM (Functional annotation of mouse)

The initial attempts of peeking into the transcriptomes mainly relied on mapping the expressed sequence tags (ESTs) (Schuler G.D. at al., 1996; Liang et al., 2000), which had a limitation of representing only a small portion of the complete cDNA (complementary DNA sequence synthesized from RNA molecule) fragments. Later in 2000, the **FANTOM** international consortium was established and led by RIKEN (The Institute of Physical and Chemical Research, Japan) in parallel to the HGP, aiming to annotate and study the transcriptional landscape of the mouse genome using full-length cDNA clones generated by the RIKEN Mouse Gene Encyclopedia Project. The major challenges faced in the cDNA identification

were - firstly, the known varied expression levels of RNA in a typical cell, which could range from very low to abundant for subsets of RNA molecules and secondly, the close resemblance of the unspliced heterogeneous nuclear RNAs (hnRNAs) with that of the unspliced non-coding RNAs. To address the latter, they focused exclusively on the cDNA specific to cytoplamic RNA molecules (Carninci et al., 2002) that are processed and polyadenylated, and ignored the wealth of non-coding RNAs present in the nucleus to avoid a possible misinterpretation of non-coding RNA with hnRNAs (heterogeneous nuclear RNA) which are a bulk of pre-mRNAs (precursor mRNA) and nuclear RNA transcripts that do not end up as cytoplasmic mRNA. To address the former issue, FANTOM followed the modified cap-trapper based strategy, that included "cDNA normalization" method to homogenize the frequency of cDNA and the "subtraction" method to delete redundant cloned cDNA from libraries (Carninci et al., 2000). This allowed for the identification of 60,770 full-length novel cDNAs corresponding to 33,409 trancriptional units (TU), each representing a region of the genome which were transcribed into one or more unique RNAs. Out of the total TUs, 11,665 and 4,258 corresponded to novel non-coding RNAs and mRNAs respectively, implying that the non-coding RNAs constitute a significant fraction of the mouse transcriptome. In addition to this, several evidence revealing the similarity between the coding and non-coding transcripts were also identified. For example, non-coding RNA transcripts showed similar poly-adenylation signals as that of coding RNAs, suggesting that they both are the product of RNA Pol II-mediated transcription processing (Hirose & Manley, 1998), and nuclear export. Furthermore, a subset of non-coding transcripts were identified to show alternative splice evidence similar to that of protein coding transcripts (Okazaki Y. et al., 2002), Altogether, these observations provided an initial glimpse of the eukaryotic transcriptome, where the non-coding RNAs were very recurrent.

The subsequent FANTOM3 project implemented new advanced techniques such as "cap analysis gene expression" (CAGE), developed for the identification of the transcription start sites and differences in the expression levels reflected due to different promoter usage (Shiraki et al., 2003), and "gene identification signature" (GIS), developed to accurately identify the 5' and 3' end signatures of the cDNA using the paired-end ditags (PETs) that could reliably be mapped to genomic sequences (Ng et al., 2005). These new technologies allowed FANTOM3 to identify a total of 102,281 unique transcript sequences, of which 34,030 were annotated as non-coding. Many of the transcript diversities were also identified due to alternative poly-adenylation signals (Carninci et al., 2005).

The large-scale transcriptome profiling of mouse by the FANTOM projects also revealed evidence for widespread transcription from both the strands of DNA, resulting into the coordinated regulation of the RNA transcribed from a given genomic locus. The **antisense transcription** (i.e. the process of transcription of an RNA from the antisense strand with respect to the overlapping coding gene) can regulate the expression of the gene in sense strand, for example, the imprinting of the *Igf2r* gene locus is due to the transcription of an antisense transcript named *Airn* (Lyle et al., 2000). Alternatively, the RNA transcribed from the antisense strand could potentially hybridize with the mRNA transcribed from the sense strand of the same locus forming a sense-antisense pair (S/AS), where the antisense transcript could either be coding or non-coding. (Kiyosawa et al., 2003; Okazaki Y. et al., 2002). Microarray based co-expression studies of such S/AS pairs revealed a complex and tissue-specific regulation. Additionally, the study of possible interactions between the pairs revealed that they could either be concordantly or discordantly regulated (Katayama, Tomaru, & Kasukawa, 2005). The S/AS pair of transcripts are also found among other organisms, as reported in

several previous and subsequent studies to FANTOM, for example, in human (Yelin et al., 2003), *Drosophila* (Misra et al., 2002) and *Arabidopsis thaliana* (Yamada et al., 2011). Hence, the S/AS transcript pairs represent an added regulatory feature of the eukaryotic transcriptomes. Altogether, the efforts of FANTOM and RIKEN produced an advanced and sensitive technique to explore the transcriptomes, using which they could identify the non-coding sequences in the genomes are prominently transcribed into novel non-coding RNA products with putative functional roles.

1.1.3.2. ENCODE (ENCyclopedia Of DNA Elements)

Alongside of the FANTOM, the **ENCODE** Project was launched in September 2003 by US National Human Genome Research Institute (NHGRI) which aimed to catalog all the functional elements in the human genome (Feingold et al., 2004). In the pilot phase ENCODE focused for the identification of transcripts features and functional elements within the 30 Mb (1%) of human genome sequences by implementing the following three methods - 1) hybridization of the RNA to tilling microarrays, a technique first designed and deployed by *Shoemaker et al., 2001* and *Kapranov et al., 2002,* 2) identification of 5' and 3' ends of the transcripts using CAGE and GIS (PET-tagging) techniques, previously developed and used by FANTOM and 3) lastly, integrated assembly and annotation of available cDNA and EST sequences involving computational, manual, and experimental approaches in-line with GENCODE annotation pipeline, a sub-project of ENCODE dedicated to the annotation of the transcribed features in the genome (Harrow et al., 2006).

The usage of tilling microarray technique by ENCODE demonstrated for the first time, the pervasive transcription in the mammalian genome, along with the identification of a large

number of novel sites of active gene expression that were previously not annotated either by the computational gene prediction algorithms or identified in massive collection of sequenced cDNAs, a strategy previously followed by FANTOM. The evidence of pervasive transcription included the transcripts that linked distal regions to establish a protein-coding loci. Many of the identified transcripts were also non-coding, a large fraction of which were identified to be in overlap with the protein-coding loci. A catalog of TSS for the transcripts were also predicted by the GENCODE annotation pipeline and the combination of transcripts identified using CAGE and PET-tagging techniques. This yielded numerous previously unrecognized TSSs that correlated with the DNaseI hypersensitive sites (DHS), active histone marks, transcript density and transcription factors such as E2F1, E2F4 and MYC (Birney et al., 2007).

The advanced techniques, methods and computational approaches experimented in the ENCODE pilot phase along with the other methods developed since 2007 such as RNA-seq for mapping the RNA transcribed regions, mass spectrometry to validate protein-coding regions, ChIP-seq and Dnase-seq for transcription factor binding sites, FAIRE-seq (Formaldehyde assisted isolation of regulatory elements), histone ChIP-seq and Mnase-seq (micrococcal nuclease) for the identification of chromatin structure and RRBS assay (Reduced representation bisulphite sequencing) for tracking DNA methylation sites, were subsequently implemented in the second phase of the project called as the ENCODE production phase. The ENCODE production phase started in 2007 with the aim to identify all the functional elements in the entire human genome (The ENCODE Project Consortium, 2012).

In parallel to the ENCODE project, the mouse ENCODE (Mouse Encode Consortium, 2012) and the modENCODE (ENCODE for model organism) consortium focusing on *D.*

melanogaster (Roy et al., 2010) and *C. elegans* (Gerstein et al., 2010) mapped the transcription, DNase1 hypersensitivity, transcription factor binding, and chromatin modifications throughout the genomes in diverse cell, tissue types and embryos *(C. elegans).* This allowed a comparative analysis between distinct species, revealing wide range of evolutionary forces acting on genes and their regulatory regions (Boyle et al., 2014; Gerstein et al., 2014; J. W. K. Ho et al., 2014).

One of the major breakthroughs of the ENCODE project was the identification and analysis of the most comprehensive manually curated large catalog of human long non-coding RNAs which were annotated and classified by the GENCODE consortium into transcript and gene biotypes (http://www.gencodegenes.org/gencode_biotypes.html), mainly based on their coding potential and the location with respect to protein-coding genes (Harrow et al., 2012).

1.1.3.3. GENCODE

The GENCODE consortium has played a crucial role in the accomplishments of ENCODE by cataloging a high quality functional annotations of the identified features. The GENCODE annotation pipeline is the combination of manual gene annotation from the HAVANA (Human and Vertebrate Analysis and Annotation) group (HAVANA - Wellcome Trust Sanger institute) and automatic gene annotation from Ensembl (Flicek et al., 2011). The combination of both manual and automated gene annotation method makes GENCODE a highly reliable resource of annotated features for the human and mouse transcriptomes, hence the GENCODE is also the primary source of annotated transcript and gene features used in my study.

1.1.3.3.1. HAVANA

The HAVANA team is based on the Wellcome Trust Sanger Institute which largely focuses on the manual annotation of the human, mouse and zebrafish genomes because manual annotation method is considered more reliable than automated annotations particularly in scrutinizing splice variation, pseudogenes, conserved gene families, duplications, non-coding genes and lncRNAs. The data used for the annotations are the combination of RNA-seq data, chromatin-state maps and computational predictions. HAVANA's manual annotations are supported and analyzed using the modified Ensembl pipeline called as the "Otter annotation system", which allows the incorporation of extra textual data necessary in support of manual annotations, for example, a descriptive text to attribute functionality to an existing gene structures etc (Searle S. M. et al., 2004; Loveland et al, 2012). The manual annotations from HAVANA are also supported by several quality-check systems developed by the GENCODE consortium such as "AnnoTrack" (Kokocinski, Harrow, & Hubbard, 2010) which can identify potential missing or incorrect manual annotations including missing loci, missing alternative isoforms or incorrect biotypes. To maintain a high quality manual annotations, GENCODE consortium also offers experimental validation pipelines based on RNA sequencing and RACE (Rapid amplification of 5′ cDNA ends) methods (Searle S. et al., 2010). The annotations from HAVANA can be freely accessed through VEGA (The Vertebrate Genome Annotation Database), Ensembl or UCSC genome browsers.

1.1.3.3.2. Ensembl

The Ensembl project was initiated in 1999 with a joint collaboration between the EBI (European Bioinformatic Institute) and the Sanger Institute to provide an automated annotation and visualization system for genomic sequences by integrating biological data and making

them publicly available via web portals. Today Ensembl is one of three main web portals that are dedicated to annotate and display the genome-scale data, other two being the UCSC (Karolchik, 2004; Rosenbloom et al., 2015) and NCBI (NCBI Resource Coordinators*, 2015; Wheeler, 2004). Along with the annotation of the human genome, the Ensembl annotation pipeline is under constant evolution and has been successfully delivering the high quality annotations of functional elements for mouse, rat, zebrafish, fly, worm, and fugu genomes. As per Ensembl release 77, it supported 69 species with complete genome annotations on the main website (http://www.ensembl.org). Ensembl has also under gone rapid expansion to incorporate genomic annotations for invertebrates in separate websites for bacteria, protists, fungi, plants and metazoa which are organized together in the Ensembl Genomes resource (http://ensemblgenomes.org), launched in 2009 (Kersey et al., 2012).

Ensembl automatic gene annotation pipeline is composed of Perl APIs (application programming interface) and the core (Ensembl MySQL) database. For Ensembl gene build, the Perl API facilitates modular execution of programs such as Genscan (Burge & Karlin, 1997) for gene prediction, RepeatMasker (Smit & Green, 2015) for the identification of interspersed repeats and low complexity DNA sequences, tRNAscan-SE for the detection of tRNA genes (Lowe & Eddy, 1997) and BLAST for the homology searches (Altschul et al., 1990). The results of these executions are stored into a MySQL database and displayed in the Ensembl websites (Potter et al., 2004). In addition, Ensembl actively collaborates with RefSeq group at NCBI, the HAVANA group at the Sanger Institute, the UCSC genome group, to establish the set of protein-coding gene structures which are stable and possibly in agreement among all the groups, this is also called as the CCDS (Consensus Coding Sequence) project which made its first release in 2005 (Pruitt et al., 2009) and is under constant updates since

then (Birney, 2006). Besides, Ensembl displays protein related information from the UniProt and Pfam databases, established by EBI, SIB (Swiss Institute of Bioinformatics), PIR (Protein Information Resource) which are also the central resources for comprehensive catalog of protein sequence and their functional annotations (Finn et al., 2014; The UniProt Consortium, 2007).

Additionally, Ensembl also integrates information related to genetic variations from the resources such as dbSNP (Sherry et al., 2001) and DGV (MacDonald et al., 2014) databases that catalogs SNPs (single nucleotide polymorphism) and CNVs (copy number variation) identified by the International HapMap project that aimed to develop a haplotype map (HapMap) of the human genome (The International HapMap Consortium, 2003) and the 1000 Genome project that aimed to produce a high-resolution map of SNPs as well-as CNVs by sequencing the genomes of 1000 individuals (The International HapMap Consortium, 2003; The 1000 Genomes Project Consortium, 2010) respectively. The HapMap project, identified more than 100 regions of the genome containing genetic variants affecting human health, disease and response to drugs and environmental factors, whereas a step beyond, the 1000 Genome project reconstructed the genomes of 2,504 individuals from 26 populations and characterized a broad spectrum of genetic variation. The project catalogued a total of over 88 million variants, more than 99 percent of which are identified as SNP variants that occurred with a frequency of at least 1 percent in the populations studied (The 1000 Genomes Project Consortium, 2015). The historically significant effort made in the 1000 genomes project also produced an integrated map of structural variations in human genomes. These finding together adds up to our understanding of the patterns of variation in individual's genomes and provide a

foundation for gaining greater insights into the genomics of human disease (Sudmant et al., 2015).

Comparative genomics is an integral part of Ensembl, for which it has developed the Ensembl Compara multi-species database which stores the results of the cross-species comparisons and analysis that includes genome alignments, syntenic regions, genome conservation, ncRNA trees, protein trees and gene trees from where the orthologues and paralogues genes are inferred.

Ensembl also provides a list of tools and services to facilitate the users to access the vast resource of data in an efficient way, for example, Ensembl offers the BioMart services as the primary data-mining tool that continues to be updated with each Ensembl release. The BioMart services are accessible by Ensembl website, alternatively there are other programmatic access available for example, BioMart's Perl, REST (Representational State Transfer) APIs (Yates et al., 2015) and the popular Bioconductor biomaRt package, that integrates BioMart data resources for data access and analysis in R programming platform. The biomaRt package enables an easy and up-to-date download of genomic data, such as gene and gene product identifiers, gene symbols, chromosomal coordinates, Gene Ontology annotations and sequences etc, by executing direct SQL queries to the Ensembl database. The biomaRt also provides an integrative powerful environment for biological data mining and analysis (Durinck et al., 2005).

For the automated annotation of non-coding RNA, Ensembl aligns the genomic sequences against RFAM (RNA family database) using BLASTN. RFAM is a comprehensive collection

of non-coding RNA (ncRNA) families that uses stochastic context-free grammars (SCFGs) (Griffiths-Jones, 2004) based identification of ncRNA families along with the combination of secondary structure and primary sequence profile of multiple sequence alignments. The miRNAs (micro-RNAs) included in the Ensembl annotations are predicted based on the alignment of genomic sequence slices against miRBase (database for micro RNAs) (Griffiths-Jones, 2006). Additionally, the cDNA alignments and chromatin-state maps from the Ensembl regulatory build (Zerbino et al., 2015) are together used to predict lncRNAs for human and mouse. This includes the identification of chromatin methylation marks across the genomic sequences (H3K4me3 and H3K36me3) outside the known protein-coding loci from the Ensembl annotation, followed by capturing cDNAs overlapping to these methylation marks that are accounted as candidate lncRNAs. Finally, the identified candidate lncRNAs are evaluated for the presence of the Pfam protein domains and substantial open reading frames (ORF). The ones carrying these features are rejected, whereas the remaining candidate lncRNAs are classified as the lncRNA genes set.

The genes annotated by Ensembl and HAVANA are merged together to produce the high quality GENCODE annotations. For the process of "gene merge", Ensembl has developed a module called "HavanaAdder", prior to the execution of which manually curated Havana gene models are passed through the Ensembl health-checking system for the identification of the inconsistencies in the annotations (Harrow et al., 2012). Next to the gene merge process the GENCODE gene features are classified into one of three broad locus level biotypes namely - protein-coding gene, long non-coding RNA (lncRNA) gene, or pseudogene based on the evidence of transcription and/or protein from the supporting source of annotation. For transcripts belonging to each of these locus biotypes, a more detailed transcript level biotypes

are assigned. For example, *Antisense* biotype for the lncRNA transcripts overlapping to the genomic coverage of one or more coding loci on the opposite strand, and *Protein_coding* biotype for the transcripts containing CDS (coding sequences) (Derrien et al., 2012; Harrow et al., 2012). The comprehensive list of all biotypes along with the definitions used in the GENCODE annotation can be found at the following webpage: http://www.gencodegenes.org/gencode_biotypes.html.

In sum, the above mentioned large-scale scale efforts have remarkably changed our initial perspectives regarding the genomes and the complexity of organisms by revealing the underlying convolution of genomic sequences and the transcriptomes with incessant discoveries.

1.1.4. The paradigm shift: Pervasive transcription in eyukaryotic genome

We have seen that the large-scale efforts forged in follow-up to the initial release of human genome sequence have revolutionized the world of genomics and led to the drastic development of technologies and methods used in the studies. The initial large scale cDNA sequencing approach and the implementation of CAGE, GIS techniques in mouse as a part of FANTOM (Carninci et al., 2005; Okazaki Y. et al., 2002), along with the application of tilling microarray techniques in human as a part of ENCODE, revealed the evidence for pervasive transcription in the mammalian genome (Kapranov et al., 2002; Kapranov et al., 2007) with a majority of the transcriptional output corresponding to the non-coding sequences of the genome that were previously considered as "Junk DNA". This arose the question, if the "Junk DNA are functional and contribute to the organismal complexity?". Some argued they are functionally active based on the initial known examples of functional non-coding RNAs such

as H19, Xist, and *Airn* in gene imprinting and X chromosome inactivation, other cellular processes (Bartolomei, Zemel, & Tilghman, 1991; Brockdorff et al., 1992; Lyle et al., 2000) and more globally in control of genetic networks (reviewed by Mattick & Gagen, 2001). Whereas, others argued the observed pervasive transcription were merely "trancriptional noise", because of their relative low expression levels in contrast to the coding RNAs and the lack in demonstration of their functionality (Ebisuya et al, 2008; Struhl, 2007; van Bakel et al., 2010; J. Wang et al., 2004). The disagreement of ideas and the limitations of the analytical techniques to determine the exact functional activities of many lncRNAs, created a sense of mystery across the scientific community leading to their definition as genomic "dark matter", in a manner analogous to the "dark matter" of the universe whose perception is difficult nonetheless its existence is known and is open for experimentations.

There are several line of studies that did not agree with the idea of functional lncRNAs. Some of such studies include, the RNA-seq data based claim by *van Bakel et al. 2010,* according to whom, the observed novel transcripts could just be a previously undetected extension of a known coding genes (van Bakel et al., 2010). However, based on a comparative study done by Clark et al., 2011, this claim was explained inappropriate due to the lack of sequencing depth and poor transcript assembly. Similarly, the large-scale identification of intergenic transcripts, including non-coding RNAs, antisense transcripts (Katayama, Tomaru, & Kasukawa, 2005) and the transcripts originating from alternative transcription start sites of known genes (Carninci et al., 2006) were also interpreted as transcriptional noise due to the described "ripples" of transcription extending from protein coding-genes (Ebisuya et al, 2008). In addition, based on bioinformatic analysis on the conservation of sequences, *Wang et al. 2004,* raised the questions against the idea that majority of 33,407 putative full-length cDNAs

identified by FANTOM in mouse were non-coding with putative functionalities (Okazaki Y. et al., 2002), as they identified a low level of conservation of non-coding sequence which were no greater than that observed for intergenic sequences (Wang et al., 2004). However, the control set used in *Wang et al.'s* study mainly contained only few already known functional non-coding RNAs (Okazaki et al., 2004). Currently, It is widely accepted that the functional non-coding RNAs are in general less conserved than protein-coding sequence, for example, the well known Xist lncRNA show low homology (60%) between mammalian species, despite retaining an identical function of X-chromosome inactivation.

1.1.4.1. Examples of functionally active non-coding RNAs

The evidence for the detected novel non-coding transcripts not being transcriptional artifacts are strong. For example, the initial detailed study on the transcriptome for human chromosome 21 and 22 using high-density oligonucleotide arrays, gave a magnified view by revealing 49% of the observed transcription were outside of any known annotation, and that these novel transcripts appeared to be more cell-line specific than well known genes, although they showed lower and less variation in expression. The lower variation in the novel transcripts were likely a result of their overall lower expression levels that emerged as a characteristic feature of non-coding transcripts based on later studies (Kampa et al., 2004; Mercer et al., 2008; Derrien et al., 2012). The mapping of transcription factors along chromosomes 21 and 22 revealed that 22% of the transcription factor binding sites (TFBS) to be present at 5' end of coding genes, whereas 36% were found at the immediate 3' end of the known coding genes, significantly correlating with the presence of a mapped non-coding RNAs. In addition, the co-regulation of overlapping protein-coding and non-coding RNAs were seen to occur significantly more often than random, suggesting that non-coding RNAs are functionally

transcribed (Cawley et al., 2004). Furthermore, *Ponjavic et al. 2007,* compared 3,122 mouse full-length lncRNAs and their promoters identified by FANTOM (Carninci et al., 2005; Okazaki Y et al., 2002) against human and rat orthologous and found purifying selection acting on the promoters, primary sequence, and consensus splice site motifs, implying they are functinoally active (Ponjavic, Ponting, & Lunter, 2007). The evidence of pervasive and non-coding transcription attracted large groups of scientific community who were interested in understanding if the non-coding RNAs are transcribed for specific functions.

Apart from well known examples of functional non-coding RNAs already mentioned, a number of studies consequently reported for the new functional roles for transcribed non-coding RNAs. One such study described the necessity of non-coding RNA *SRG1* in the repression of *SER3* coding gene in yeast *(S. cerevisiae)* through transcription-interference mechanism, where the transcription of SRG1 across the SER3 promoter interferes with the binding of transcription activators necessary for SER3 transcription. (Martens, Laprade, & Winston, 2004). Non-coding RNAs are also reported to be influencing transcription of coding genes in an RNA mediated manner, for example, the human gene encoding for dihydrofolate reductase (DHFR) is known to have alternative promoters called as the major and minor promoters. Almost 99% of DHFR RNA transcription are known to originate from the major promoter, whereas the the minor promoter which is present upstream to the major promoter is known to initiate the transcription of a non-coding RNA. The transcribed non-coding RNA from minor promoter is found to be involved in, direct interactions with the transcription factor IIB and the *major* promoter, to form a stable complex eventually leading to the dissociation of pre-initiation complex and promoter-specific transcriptional repression at major promoter (Martianov et al., 2007). Subsequently, a long non-coding RNA called as *HOTAIR*

was identified to epigenetically repress the expression of *HOXD* gene cluster in trans across 40 kilobases by recruiting the polycomb repressive complex 2 (PRC2), which is further required to repress the histone H3 lysine-27 trimethylation within the *HOXD* cluster of genes (Rinn et al., 2007).

The studies mentioned above constitute only few initial important examples of large number of evidence currently available for functional characteristics of non-coding RNAs (*reviewed by* Mercer, Dinger, & Mattick, 2009; Ponting, Oliver, & Reik, 2009). Taken together, all these studies imply that the long non-coding RNAs are far from being transcriptional "artifacts" or "noise". Instead, they are involved in important molecular functions which have been exhaustively cataloged and are endlessly growing till-date. Although there are presently several lines of evidence supporting the claim that some lncRNAs are functional through RNA-mediated mechanisms, the identification and extent of lncRNAs involvement in active transcriptional activities and underlying biological implications still remains unresolved for many of them. However, now at least it is clear that once which seemed as "Junk DNA", later regarded as the genomic "Dark Matter", are now being uncovered as the non-coding RNAs which are functional in nature and rendering several kinds of important molecular roles.

1.2. Non-coding RNAs and the characterization of long non coding RNAs

Before going any further into the functional classes of lncRNAs, it is important to discuss how are they characterized and where they stand among different families of non-coding RNAs. Non-coding RNA (ncRNAs) as a whole represents the set of RNA molecules which are transcribed but do not encode for proteins. Some of the earliest RNA molecules to be categorized as the ncRNAs were, transfer RNAs (tRNAs), ribosomal RNA (rRNAs), small

nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) which together are also referred as house-keeping or structural RNAs, because of their generic cellular functions. The functions of tRNAs and rRNAs are well established in the mRNA translation process, snRNAs are involved in splicing mechanisms (Grabowski, Seiler, & Sharp, 1985), whereas snoRNAs are found functional in the modifications of rRNA molecules (Lafontaine & Tollervey, 1998; Makarova & Kramerov, 2011). For a long period of time the non-coding RNAs were thought to be a series of accessories that are needed to process genes to make proteins, with tRNAs serving to decode the codons in mRNA and provide amino acids in the order they are needed for insertion into the growing polypeptide chains, rRNAs constituting the essential components of the ribosomes, the complex ribonucleoprotein factories of protein synthesis and other ubiquitous ncRNAs such as snRNA and snoRNA to function higher up the pathway to ensure correct processing of mRNA, tRNA and rRNA precursors. However, the identification of the first functional lncRNA, H19 in mouse, (Brannan et al., 1990) which is involved in parental gene imprinting (Bartolomei, Zemel, & Tilghman, 1991) indicated the possible existence of diverse functionalities of ncRNAs and their types. Subsequently, we have witnessed for the dramatic increase in the catalog of functional lncRNAs having implications in gene regulatory activities, complex biological processes and disease (reviewed in Amaral & Mattick, 2008; Dinger, Gascoigne, & Mattick, 2011; Goff & Rinn, 2015; Mattick, 2006; Ponting, Oliver, & Reik, 2009).

Based on functions, the ncRNAs can be broadly classified as structural RNAs and regulatory RNAs. The regulatory RNAs can be further classified as small ncRNAs and long ncRNAs based on their lengths. MicroRNAs (miRNAs), Small-interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs) together constitute the small ncRNA class (sncRNA). miRNAs

(21-25 nt) and siRNAs (20-25 nt) are profoundly studied for their post transcriptional regulatory activities (Amaral & Mattick, 2008; van Wolfswinkel & Ketting, 2010). Whereas piRNAs, (26-31 nt) which are slightly longer in length compared to miRNAs and siRNAs are found to be expressed in spermatogenic cells in the testis of mammals. They are well know for their retrotransposon silencing activity (Kim, 2006).

In contrast to small ncRNAs, the lncRNAs embodies a major class of ncRNA. LncRNAs are the transcripts longer than 200 nt (nucleotides) that lack an open reading frame (ORF) (Derrien et al., 2012; Dinger et al., 2011; Harrow et al., 2012; Ponting et al., 2009) and are known to be transcribed either in antisense to protein-coding genes or as intergenic or intronic (Amaral & Mattick, 2008; Kapranov et al., 2007; Lehner et al., 2002). The lncRNA expression is developmentally regulated and they are known to exhibit high tissue or cell-type specificity (Wilusz, Sunwoo, & Spector, 2009). LncRNAs also share a high degree of structural similarity with coding mRNA transcripts, indeed they are transcribed by RNA polymerase II and might acquire poly-adenylation in their 3' terminal ends (Tuck & Tollervey, 2013). However, in contrast to mRNAs and other structural RNAs, lncRNAs are localized in the nucleus, whereas only a small fraction of lncRNAs are found in both cytoplasm and nucleus or specifically distributed in cytoplasm (Kapranov et al., 2007; Mercer et al., 2009; Nie et al., 2012).

1.2.1. Evolutionary conservation of lncRNAs

The evolutionary conservation of lncRNAs appears to be very less pronounced compared to the protein-coding genes and other small ncRNAs such as miRNAs and snoRNAs (Pang, Frith, & Mattick, 2006; Qu & Adelson, 2012). As the evolutionary conservation is a widely used criteria to predict functional significance of newly discovered genes, the general lack of

conservation associated to lncRNAs has been a major point of debate. LncRNAs are found to most likely exhibit conservation of very short stretches of sequences which are believed to maintain their functional domains and structures (Pang, Frith, & Mattick, 2006), for example, the well studied lncRNAs, such as Xist, with a well established functional role in X-chromosome dosage compensation are known to have poorly conserved sequences (Duret et al., 2006), Nevertheless, they contain an important functional domain of 1.6 Kb region (out of a total transcript length of about 17 Kb) known as RepA, which is conserved across mammals and comprises 7.5 tandem repeats of 28 nt sequence. The conserved domain is known to fold into the conserved stem-loop structure that is involved in the recruitment of PRC2 protein complex thus serving the function of Xist in the X chromosome inactivation (Wutz, Rasmussen & Jaenisch, 2002; Zhao et al., 2008). Similar to Xist, the secondary structure of HOTAIR is also widely conserved across mammals (He, Liu, & Zhu, 2011). These examples imply that the lncRNAs have a very different grammar in respect to coding RNAs, as they maintain well conserved functions across species despite of their very little sequence conservation.

Currently, the elucidation of the conservation of lncRNA relies on four different main characteristics – 1) the sequence, 2) the structure, 3) the function and 4) the genomic position or more specifically the expression from syntenic loci (Diederichs, 2014) with respect to the flanking protein-coding genes. Recently, several line of studies considering human, mouse and zebrafish have shown that protein coding genes lying near a lncRNA gene have a higher probability to have their orthologs flanking an lncRNA genes (Basu, Müller, & Sanges, 2013; Necsulea et al., 2014; Ulitsky et al., 2012).

1.3. Large-scale identification of long non-coding RNAs

We have seen the GENCODE annotation pipeline produced the most comprehensive set of lncRNAs for human (GENCODEv7) (Derrien et al., 2012). However, apart from GENCODE many independent groups also contributed to the identification and annotations of lncRNAs. For example, *Cabili et al. 2011,* developed a computational approach for the comprehensive annotation of lncRNAs *(intergenic lncRNAs)* by taking advantage of existing RNA-seq data with available annotations from different sources such as GENCODEv4, RefSeq and reconstructing the transcriptomes using two assembler softwares called Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010). To cross validate the assembled transcripts against the annotated source transcriptomes and to determine the unique set of isoforms for each transcript locus, they used Cuffcompare (Trapnell et al., 2010). However, the two main challenges in the annotation of lncRNA gene loci are – 1) to distinguish the lowly expressed lncRNAs from other lowly expressed single exons and 2) distinguishing novel transcripts encoding proteins or short peptides from bona fide non-coding ones. To address the former challenge *Cabili et al* removed all unreliable lowly expressed transcripts using a read coverage threshold and focused only on multiexonic transcripts, whereas to address the later, they used phylogenetic codon substitution frequency (PhyloCSF) (Lin, Jungreis, & Kellis, 2011) to remove any putative ORFs that were evolutionarily constrained to preserve synonymous amino acid content. Additionally, they also scanned each transcript in all three reading frames to exclude transcripts that encode any of the protein domains cataloged in the protein family database Pfam. Using the above methodology *Cabili et al.,* could identify a reference catalog of 8195 human lncRNAs, 58% of which were novel and were identified for the first time exclusively using RNA-seq data (Cabili et al., 2011). The identified set of lncRNA transcripts also demonstrated a high tissue specific expression with respect to coding genes.

Using the RNA-seq experiments during zebrafish embryogenesis *Pauli et al. 2012,* could assemble the transcripts derived from known zebrafish RefSeq annotated genes and Ensembl gene models (Flicek et al., 2011). They additionally identified 1133 new lncRNAs that included lincRNAs, intron overlapping lncRNAs, exonic antisense overlapping lncRNAs, and siRNA precursor lncRNAs. The identified set of zebrafish lncRNAs showed similar characteristics to that of mammalian lncRNAs, for example, the tissue-specific expression pattern, relative short length, low overall expression level and a smaller number of exons. Furthermore the identified novel lncRNAs showed a narrow window of expression compared to protein-coding mRNAs and were specially enriched during early stage embryos pointing towards their specific functionalities during the developmental process (Pauli et al., 2012).

1.3.1. Challenges in the identification of lncRNAs

One of the major challenges faced in identification of lncRNAs in the above mentioned studies is the ability to distinguish between the coding and non-coding RNAs. Given that the lncRNAs and mRNAs are synthesized by RNA polymerase II, they share a high degree of resemblance in many of their features such as 5' capped structure, exon/intron length, 3' poly-adenylation and histone modification (Guttman et al., 2009; Tuck & Tollervey, 2013). Additionally, the relative low expression levels of lncRNA and the general low sequence conservation makes the identification of lncRNAs even more challenging (Dinger et al., 2008). Hence, the reliable identification of lncRNAs mainly depends upon the determination of the coding potential of their sequences.

1.3.2. Strategies of the lncRNA identification

Currently there are several published methods for the determination of the coding potential of the RNA molecules, each based on specific strategy for the identification of true lncRNAs. The CONC ("coding or non-coding") is one the early such method developed by FANTOM. CONC classifies the transcripts as coding or non-coding based on the features they would have if they were coding for proteins, for example, peptide length, amino acid composition, predicted secondary structure content, number of homologs from database searches, and alignment entropy (Liu, Gough, & Rost, 2006). Other methods include, the CPC (coding potential calculator) that searches transcripts for putative ORFs and homologies in UniRef database that offers a complete coverage of protein sequence space at several resolutions while hiding redundant sequences (Kong et al., 2007; Suzek et al., 2007). The PORTRAIT is one of the similar tool which rely on the *ab initio* features of the transcripts such as the ORF length and nucleotide composition for the classification of lncRNAs (Arrial, Togawa, & Brigido, 2009). The working principle of these tools are based on the support vector machine (SVM), machine learning approach which is a widely used classification tool in bioinformatics analysis.

Apart from the above mentioned tools there are several alignment-based approaches too, for the identification of lncRNAs. The PhyloCSF is one such tool that analyzes a multispecies nucleotide sequence alignment to determine whether the transcript sequence is likely to represent a conserved protein-coding region based on the codon substitution frequencies across multiple species and scores for conserved ORFs (Lin M. F. et al., 2011). The RNAcode is the another tool for the robust discrimination of coding and non-coding regions which is

based on the alignment of homologous nucleotide sequences. It predicts local regions of high coding potential together with an estimate of statistical significance (Washietl et al., 2011).

On the other hand the CPAT (Coding potential assessment tool) uses a logistic regression model for the identification of coding potential considering ORF size, ORF coverage, Fickett TESTCODE score (a metric based on the nucleotide composition and codon usage bias) and hexamer usage bias (Wang L. et al., 2013). Away from the computational prediction algorithms, a recently developed "Ribosome Profiling" technique (Ingolia, 2009) has opened up an efficient experimental approach to identify if RNA molecules are being actively translated. The technique relies on the identification and the deep sequencing of the "ribosomal footptints" *(portion of RNA molecule to which the ribosome is attached)* hence targets specifically the RNA sequences protected by ribosome during the process of translation. The determination of the ribosome occupancy for RNA molecules along with a metric called the ribosome release score (RRS) which indicates the termination of translation at the end of an ORF by tracking the ribosome's encounter with a bona fide stop codon, indicates an accurately distinction between a protein-coding and non-coding transcripts (Guttman et al., 2013). More recently, *Cenik et al, 2015,* explained the application of ribosome profiling in the computation of accurate translation efficiency of RNA molecules using a linear modeling approach considering RNA expression levels along with the ribosome occupancy of the RNA (Cenik et al., 2015).

1.3.3. lncRNA specific databases

The large-scale identified lncRNAs are cataloged into multiple comprehensive databases, Ensembl being the central database with current release version 82 with GENCODE v23 for

human and M7 for mouse. The other sources of lncRNA catalog includes NONCODE (www.bioinfo.org/noncode) which integrates the lncRNAs from literature, RefSeq, Ensembl and is under constant updates (Liu. C et al., 2005; Xie et al., 2014). The lncRNAdb, (www.lncrnadb.org) (Quek et al., 2015) that integrates the Illumina Body Atlas expression profiles and is a member of RNA-central (Bateman et al., 2011). lncRNAdb is under compliance with the International Nucleotide Sequence Database Collaboration (Karsch-Mizrachi, Nakamura, & Cochrane, 2012). Similarly, LNCipedia (www.lncipedia.org) (Volders et al., 2013) contains lncRNAs identified by various methods along with the secondary structure information, protein coding potential and microRNA binding sites. In addition LNCipedia also integrates a strategy for detecting potentially coding lncRNAs by automatically re-analyzing the large body of publicly available mass spectrometry data in the PRIDE database (Martens et al., 2005). lncRNAMap, (http://lncrnamap.mbc.nctu.edu.tw/php/) catalogs human lncRNAs which are known as the precursor for siRNAs and which can act as decoys for miRNAs (Chan, Huang, & Chang, 2014). PlncDB, (http://chualab.rockefeller) is a plant specific lncRNA database (Jin et al., 2013). LncRNAtor, (http://lncrnator.ewha.ac.kr) (Park. C, et al., 2014) compiles the lncRNAs specifically for human, mouse, zebrafish, fruit fly, worm and yeast. LncRNAtor also integrates the RNA-seq expression data from ENCODE, modENCODE, GEO along with the evolutionarily conserved lncRNAs with correlated expression between human and six other organisms to identify functional lncRNAs. Furthermore, it stores the information of protein and lncRNA interactions by collecting and analyzing publicly available CLIP-seq or PAR-CLIP sequencing data. Other databases includes, lncRNome (*human specific;* genome.igib.res.in/lncRNome/) (Bhartiya et al., 2013), fRNAdb (www.ncrna.org/frnadb) (Mituyama et al., 2009),

1.4. Strategies to screen functional activities of lncRNAs

With the large-scale identification of lncRNAs, a number of studies also reported a huge ambit of functional activities in various cellular processes and a wide spectrum of diseases (Dinger et al., 2011; Mercer et al., 2009; Ponting et al., 2009). The screening of these functional activities mainly relies on a variety of molecular techniques, well reviewed by Goff & Rinn, 2015. Briefly, some of these important techniques include the Antisense oligonucleotides (ASO) based targeting of lncRNAs. In this technique an antisense oligonucleotides is designed to bind to the target RNA by well characterized Watson-Crick base-pairing, once the oligonucleotide is bound to the target lncRNA, it modulate the function of the target through a variety of post-binding events (Bennett & Swayze, 2010). *Sarma et al. 2010,* employed this technique to target Xist RNA, and showed showed that the targeted binding lead to the displacement of Xist RNA from the X chromosome with a fast kinetics without any observable effect on its stability. ASOs have single one strand, hence are considered adequate for their delivery to the cultured cells and animal models target. At the same time there are also some limitations associated with of ASOs, for example, ASOs requires chemical modifications to be active inside cells (Watts & Corey, 2012), and ASOs are also seen to show off-target silencing due to the off-target nucleotide-binding (Frazier, 2014). The RNAi-mediated targeted knockdown technique is another similar approach which is widely used for investigation of the loss of function and change in phenotypes by using an engineered small interfering RNA (siRNA) or sharp hairpin-shaped RNAs (shRNAs) (Paddison et al., 2002; Rao et al, 2009) that can target lncRNAs (Guttman et al., 2011). It is a more preferred technique over ASO. for cell culture experiments because it do not require any chemical modifications, as an unmodified siRNA works with high potency (Watts & Corey, 2012). Although RNA-i-mediated technique is proven to be extremely useful in the dissection of the functional role of lncRNAs, it is prone to

several off-target effects, for example a gene expression profiling study to characterize the specificity of gene silencing by siRNAs in cultured human cells revealed the instances of direct silencing of non-targeted genes that contained as few as eleven contiguous nucleotides of identity to the siRNA (Jackson et al., 2003)

Recently, a new tool based on a bacterial CRISPR-associated protein-9 nuclease (Cas9) from *Streptococcus pyogenes* has generated considerable excitement in the scientific community, which makes use of the bacterial CRISPR-Cas9 system to take control of the RNA transcription rates (Cong et al., 2013). The CRISPR-Cas9 system are also used for the targeted knock down of the lncRNAs (Ho et al., 2015). The functions of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and CRISPR-associated (Cas) genes are essential in adaptive immunity in selected bacteria and archaea which enables the organisms to respond to and eliminate the invading genetic material. These repeats were initially discovered in 1980s in *E. coli*, but their function weren't confirmed until 2007 by *Barrangou et al.*, who demonstrated that the *S. thermophilus* could acquire resistance against a bacteriophage by integrating a genome fragment of the infectious virus into its CRISPR locus. The integrated genome fragments are also called as "CRISPR spacer", which provides specificity to recognize exogenous genetic elements of virus, whereas cas9 nuclease cuts the virus genome in a manner analogous to RNA interference in eukaryotic organisms (Barrangou et al., 2007). However, without exception, the CRISPR RNA-guided nucleases (RGNs), were also identified for off-target cleavage of CRISPR-associated (Cas)9-based RGNs, where the off-target sites were found to harbor up to five mismatches (Fu et al., 2013).

The off-target activities *(discussed above)* shown by the advanced techniques used to screen functional activities can complicate the interpretation of phenotypic effects in gene-silencing experiments and can potentially lead to the false discoveries and unwanted toxicities. (Cho et al., 2014; Fu et al., 2013; Jackson et al., 2003; Lin. X et al., 2005). However, advancements are being made in order to improve the specificity of both RNAi and CRISPR-Cas9 techniques (reviewed by Barrangou et al., 2015). In parallel, other advanced methods to dissect the lncRNA functionalities are also being developed, for example, the live-cell imaging approach. The live-cell imaging system has been successfully used to monitor the activity of induced transcription of NEAT1 lncRNA (Mao et al., 2011). In sum, it is clear that the screening of lncRNA functionalities is a complex procedure and it is important to consider multiple approaches to untangle the precise mechanism of action of putative the lncRNA genes. However, the techniques mentioned above are under constant evolution and have contributed a milestone of knowledge regarding lncRNA functions.

1.5. Classification of lncRNA based on their mode of functional activities.

The systematic scrutinization of lncRNAs and their functional activities demand a standardized classification framework. Currently, the known catalog of lncRNAs can be classified in a number of ways, elaborately discussed in recent reviews (Goff & Rinn, 2015; Ma, Bajic, & Zhang, 2013). For the sake of simplicity in discussion, here the lncRNAs are broadly categorized based on their regulatory mechanisms at transcriptional and post-transcriptional levels.

1.5.1. Transcriptional regulation by lncRNAs

The lncRNAs that act as transcriptional regulators, can either act locally *(in cis)* in respect to the site of their synthesis to influence the expression of nearby genes in the same locus, or act distally *(in trans)* with respect to their genomic locus to regulate expression of several genes across chromosomes. They can act either by directly interacting with the chromatin modifying enzymes and nucleosome remodeling factors to control chromatin structures, or by interfering with the transcription of other genes.

1.5.1.1. Transcriptional interference

Transcriptional interference is a molecular event where the transcription of a gene can have a suppressive influence on the transcription of another gene present in its close proximity on the genome (Shearwin, Callen, & Egan, 2010), for example in yeast, the transcription of SER3 regulatory gene 1 (SRG1) lncRNA which overlaps with the promoter region of SER3, leads to the increase in nucleosome density at the overlapping promoter region thereby making it inaccessible for the transcriptional protein machinery leading to the transcriptional silencing of SER3 gene (Martens, Laprade, & Winston, 2004). Similarly, in mouse the transcription of

Airn lncRNA which is organized as antisense to *Igf2r* (Insulin-like growth factor 2 receptor) protein-coding gene is known for the silencing of *Igf2r* through transcriptional interference leaving behind the traces of the increased nucleosome density and DNA methylation at the promoter region of *Igf2r* (Santoro et al., 2013). However, it was demonstrated that the transcription of *Airn* alone is sufficient to suppress the transcription of *Igf2r* while the observed repressive chromatin features were simply a second layer of repression (Kornienko et al., 2013). Additionally, in mouse placenta, the imprinted *Airn* is a paternally expressed lncRNA which is also known to act in a long range silencing of cis-linked Slc22a2 and Slc22a3 (Solute Carrier Family 22 *Organic Cation Transporter* Member) genes by targeting repressive histone modifications to their imprinted promoters (Nagano et al, 2008; Sleutels, Zwart, & Barlow, 2002; Zwart et al., 2001), thereby regulating a cluster of genes in a tissue-specific manner. Similar to *Airn* and SRG1 mode of actions, a recent study on fission yeast reported a new lncRNA called "nc-tgp1", which is when transcribed, increases the nucleosome density preventing the access of transcription factors for the transcription of tgp1+ *(transporter for glycerophosphodiester 1)* gene present in the near by region (Ard, Tong, & Allshire, 2014). Apart from the mechanisms already mentioned, the transcriptional interference is also seen to occur by "transcriptional collision". An atomic force microscopy based imaging at single molecule resolution in *Escherichia coli* revealed for first time, that RNA polymerase heading towards each other could colloid during an event of convergent transcription on a linear DNA template with two convergently aligned promoters, leading to a transcriptional stall (Crampton et al., 2006).

1.5.1.2. Chromatin Remodeling

Chromatin remodeling is the process of dynamic modifications of the chromatin structure which either allows or blocks the transcription of regulatory proteins to have access to the genomic DNA and control the gene expression of genes present into specific loci. Some of the significant examples of lncRNAs taking control on the gene regulation through this mechanism include Xist, HOTAIR, HOTTIP and COLDAIR.

Xist is one of the longest known lncRNAs measuring 17 Kb in mouse and over 19 kb in humans. It has the functional activity of transcriptional inactivation of one X chromosome in female mammals (Brockdorff et al., 1992; Lee, Davidow, & Warshawsky, 1999; Pontier & Gribnau, 2011). It is transcribed from the Xist gene, which lies in the X inactivation center (Xic) within the X chromosomes. It is specifically expressed from the X chromosomes which gets inactivated. The Xist transcript recruits the polycomb repressive complex 2 proteins (PRC2) which leads to the trimethylation of lysine 27 of histone H3 (H3K27me3) based silencing of the genes in its proximity (Wutz, 2011). HOTAIR (HOX antisense intergenic RNA) is another extensively studied lncRNA known to repress the HOXD cluster of genes in humans and therefore controlling the definition of the body plan of the developing embryo. HOTAIR is a 2.2 kb *trans*-acting lncRNA, which is transcribed from the HOXC locus 40kb away from the HOXD cluster (Rinn et al., 2007). HOTAIR is know to interact with Polycomb Repressive Complex 2 (PRC2) leading to PRC2 occupancy and histone H3 lysine-27 trimethylation of HOXD locus resulting in its silencing. In contrast to HOTAIR, human HOTTIP (HOXA transcript at the distal tip) is a *cis*-acting lncRNA known to activate the transcription of its flanking genes. HOTTIP is transcribed from the 5' end of the HOXA locus and is involved in the definition of the body plan of developing embryos. HOTTIP is also

shown to act by binding WDR5 (WD repeat-containing protein 5) in the MLL (Mixed-Lineage, Leukemia) histone modifier complex thereby bringing histone H3 lysine-4 trimethylation (H3K4me3) to the promoters of its flanking genes (Wang. K. C et al., 2011). In general the mechanism from which lncRNAs deliver epigenetic modifiers to their specific target genes, while they are still attached to the elongating RNAP II is termed "tethering" and is often used to explain the positive cis-regulation by lncRNAs (Guttman & Rinn, 2012). Another example of *cis*-acting lncRNA in chormatin remodeling is the COLDAIR (Cold Assisted Intronic Non-coding RNA) lncRNA studied in plant (Arabidopsis thaliana). It is transcribed from the intron of the FLC (Flowering Locus C) protein coding gene also known as potent flower repressor and can tether repressive chromatin marks (H3K27me3) to the FLC gene in *cis* leading to its silencing thereby taking control of the flowering time during the vernalization process (Heo & Sung, 2011). These examples together provides a glimpse of our current understanding of lncRNA interaction with chromatin structures for the gene regulation process, other advancements are elaborately discussed in a recent review by (Rinn, 2014).

1.5.2. Post-transcriptional regulation by lncRNAs

With the above discussion it is noted that lncRNAs are widely implicated in the regulation of gene transcription. However, examples of gene regulation by lncRNA at post-trancriptional level are also emerging and are seemingly involved in several ways. For example, lncRNAs are known to be involved in pre-messenger RNA (pre-mRNA) splicing, mRNA stability, translation, other protein activities and even as microRNAs (miRNAs) sponges in sequence-dependent and sequence-independent manner (well reviewed in Shi et al., 2015; Yoon, Abdelmohsen, & Gorospe, 2013).

1.5.2.1. lncRNAs influencing pre-mRNA Splicing

Alternative splicing is considered as the key characteristic features of mRNA translation as well as regulation of gene functions in higher eukaryotes (Matlin, Clark, & Smith, 2005), it is modulated by snRNAs, hnRNAs and other trans-acting protein factors such as serine/arginine-rich (SR) family of nuclear phosphoproteins (SR proteins) (Grabowski et al., 1985; Long & Caceres, 2009). Interestingly, MALAT1 lncRNA was found to be interacting with the SR proteins and influencing (Tripathi et al., 2010) their distribution along with the other splicing factors such as nuclear speckle domains which represent sub-nuclear structures enriched in pre-mRNA splicing factors and are located in the inter-chromatin regions of the nucleoplasm of mammalian cells (Lamond & Spector, 2003). The depletion of MALAT1 resulted in changes of alternative splicing in specific set of endogenous pre-mRNAs and increase in dephosphorylated pool of SR proteins, as MALAT1 also regulates the cellular levels of phosphorylated form of SR proteins (Tripathi et al., 2010). Another example of lncRNA which binds directly to the splicing factor is MIAT (myocardial infarction associated transcript), also referred as Gomafu or RNCR2. MIAT lncRNA was originally identified in a particular set of

neurons in the mouse retina. Unlike protein-coding mRNAs, MIAT lncRNAs could escape nuclear export and stably accumulate within the nucleus, making a specific nuclear compartment (Tsuiji et al., 2011). A comparative genomic study reported MIAT genes from three distinct species contained a tandem repeat of the "UACUAAC" motif, which is an essential and conserved intron branch point sequence (BPS) in the budding yeast *S. cerevisiae* (Reed & Maniatis, 1988). The tandem repeat sequence in MIAT is identified to aid the binding of SF1 proteins (splicing factor 1), hence proposed to be involved in the regulation of splicing (Tsuiji et al., 2011).

1.5.2.2. lncRNAs influencing mRNA stability

Recently, gadd7 (growth-arrested DNA damage-inducible gene 7) lncRNA was found to be controlling cell-cycle progression in CHO-K1 (Chinese hamster ovary) cells by altering the mRNA stability in response to UV irradiation (Liu. X et al., 2012). The cellular stress induced by the exposure of UV in CHO-K1 cells, led to a substantial increase in the gadd7 lncRNA expression levels that could directly bind to TDP-43 (TAR DNA-binding protein), a protein which can act as an activator of the expression of Cdk6 (cyclin-dependent kinase 6) gene that in-turn is known to associate with Cyclin D and regulate G1/S transition of the cell cycle. The binding of gadd7, a DNA damage-inducible lncRNA with TDP-43 was found to disrupt the interaction of TDP-43 with Cdk6 mRNA leading to the instability of cdk6 mRNA. Another significant example of lncRNA influencing mRNA stability is Half-STAU1-binding site lncRNA (1/2-sbs lncRNA) in mammalian cells that can aid to the degradation of the target mRNAs from Staufen1 (STAU1)-mediated messenger RNA decay (SMD) pathway (Kim Y. K. et al., 2007). This involves the recruitment of UPF1 (Up-frameshift protein 1), a nonsense-mediated decay factor to the Staufen1 binding site which can be formed by imperfect base-

pairing between an Alu element in the 3' UTR of an SMD target and another Alu element in a cytoplasmic polyadenylated long non-coding RNA (lncRNA) (Gong & Maquat, 2011). Recently, STAU2, a paralog of STAU1, has also been reported to mediate SMD where both STAU1 and STAU2 interact directly with the ATP-dependent RNA helicase UPF1, enhancing its helicase activity to promote effective SMD (Park & Maquat, 2013).

Apart from the splicing regulation and influencing mRNA stability and decay of mRNA transcripts discussed above, lncRNAs are also known to act on mRNAs post-transcriptionally. For example, linc-MD1 is a muscle specific lncRNA known to play a role of decoy for two specific miRNAs, miR-206 and miR-133. These miRNAs can in turn repress the expression of MAML1 (Mastermind-Like 1) and MEF2C (Myocyte Enhancer Factor 2C) transcription factors that activate muscle-specific gene expression during the myogenic differentiation (Cesana et al., 2011). The expression of linc-MD1 leads to the subtraction of miR-206 and miR-133 permitting the MAML1 and MEF2C genes to initiate the myogenic differentiation. Another example of similar lncRNA is the BACE1-AS (BACE1-antisense), which is a highly conserved natural antisense lncRNA transcribed antisense to the BACE1 gene (beta-secretase-1), a gene known to be involved in the processing of amyloid precursor proteins (APP). In Alzheimer's disease, the BACE1-AS transcript is known to compete against the miRNA miR-485-5p and mask its binding site in the BACE1 mRNA open reading frame thereby preventing it from the miRNA induced repression (Faghihi et al., 2010).

1.5.2.3. lncRNA in translational control

Recently, lncRNAs have also been identified to be taking part in the translational control of coding mRNAs. For example, *Yoon et al.*, reported LincRNA-p21 *(intergenic lncRNA)* as a

translational modulator of JUNB (jun B proto-oncogene) and CTNNB1 (Catenin *Cadherin-Associated Protein,* Beta 1) genes in human cervical carcinoma HeLa cells (Yoon, Abdelmohsen, & Gorospe, 2013). JUNB encodes for transcription factor involved in regulating the gene activities following the primary growth factor responses, whereas CTNNB1 encodes for a protein involved in signal transduction, cell to cell adhesion and gene transcription. *Yoon et al.,* identified that the lincRNA-p21 could selectively localize to cytoplasm and exert repressive effects on the translation of JUNB and CTNNB1 genes. The activity of lincRNA-p21 was however identified to be dependent upon HuR (human antigen R) proteins which are well known to affect mRNA stability and translation by competing or cooperating with mRNA decay- promoting RBPs and miRNAs (e.g., miR-122, let-7-loaded RISC [RNA miRNA-induced silencing complex]) (Yoon et al., 2013; Kim et al., 2009). The findings of *Yoon et al.,* also suggests that HuR and let-7/Ago2 (lethal-7/argonaute RISC catalytic component 2) represses lincRNA-p21 expression cooperatively and that HuR and let-7/Ago2 binding to lincRNA-p21 are crucial for lincRNA- p21 decay.

Recently, an elegant work published by *Carrieri et al.,* shed lights on a novel post-transcriptional role for an antisense lncRNA. They reported a lncRNA antisense to the *Uchl1 coding gene (Figure 1.2a)* is able to exert up-regulation of UCHL1 protein translation, upon its transfection into mouse MN9D dopaminergic cell lines without affecting the endogenous *Uchl1* mRNA expression *(Figure 1.2b).* This lncRNA is addressed as *AS-Uchl1.* The co-transfection of *AS-Uchl1* and the murine *Uchl1* into HEK cells which do not originally express either of these transcripts, also showed an *AS-Uchl1* dose-dependent increase in the UCHL1 protein levels. This clearly indicates that the AS-Uchl1 is able to regulate UCHL1 protein translation at a post-transcriptional level *(Figure 1.2c).*

The activity of *AS-Uchl1* was identified to be under control of stress signaling pathways, as the inhibition of mTORC1 *(mechanistic target of rapamycin, complex1)* by rapamycin caused the increase in UCHL1 protein levels, which was determined by the shuttling of *AS-Uchl1* RNA from the nucleus to cytoplasm. The authors proposed that the overlap of *AS-Uchl1* with Uchl1 mRNA in cytoplasm forms a sense-mRNA/antisense-lncRNA (S/AS) pair of transcripts, where the *AS-Uchl1* can activate polysomes for the translation of overlapping Uchl1 mRNA (Carrieri et al., 2012). The activity of *AS-Uchl1* was demonstrated to mainly rely on two characteristics – 1) the S/AS 5' overlapping region and 2) the presence of *inverted* SINEB2 repeat near 3' end of the transcript. The deletion or disruption of either of them resulted in the loss of *AS-Uchl1* activity, highlighting the importance of its modular nature in the exertion of its functions. The search for similar such natural ASlncRNAs revealed *AS-Uxt* antisense to ubiquitously expressed transcript *(Uxt)* gene containing similar SINEB2 element also showed an identical regulation of UXT protein levels without affecting *Uxt* mRNA expression in mouse MN9D cells. Thus, the *AS-Uchl1* and *AS-Uxt* together represents a new functional class of natural ASlncRNAs that could activate the translational up regulation of the sense overlapping genes. This class of natural antisense lncRNAs are also referred as SINEUPs, mainly because, their activity depends upon the embedded inverted SINEB2 sequence (Zucchelli et al., 2015a).

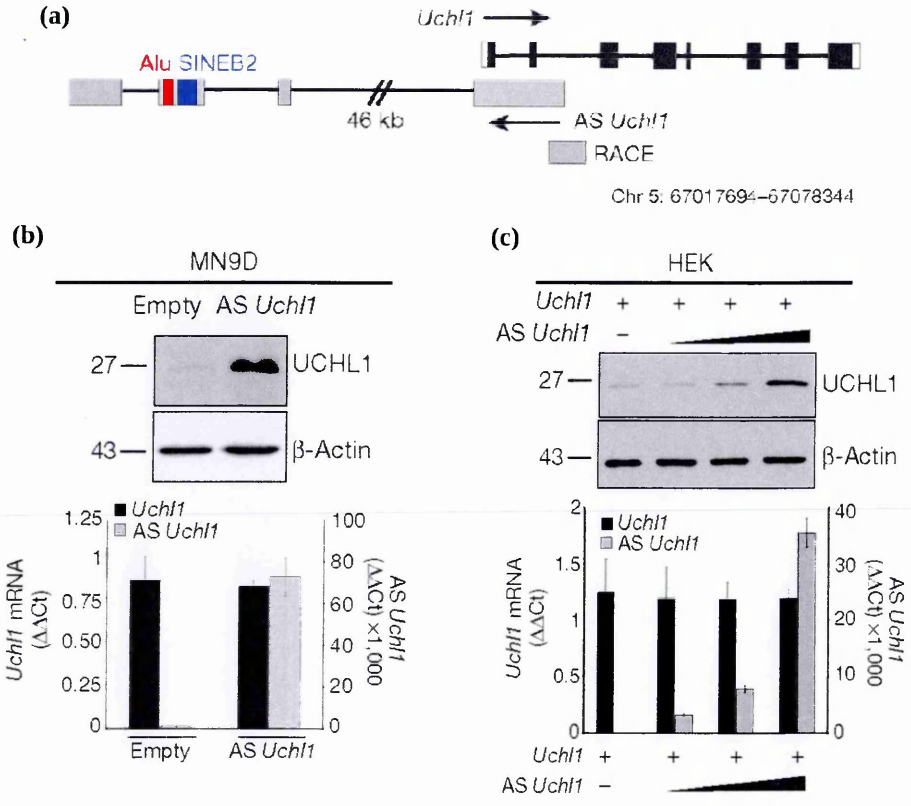


Figure 1.2 | Post-transcriptional protein up-regulatory activity of *AS-Uchl1*. (a) The genomic organization of *Uchl1/AS-Uchl1* transcripts. *Uchl1* exons are in black; 3' and 5' UTRs are in shown white; *AS-Uchl1* exons are shown in grey; repetitive elements are represented in red (Alu) and blue (SINEB2) whereas introns are indicated as lines. (b) *AS-Uchl1* transfected dopaminergic MN9D cells show increased levels of endogenous UCHL1 protein, with unchanged mRNA quantity. (c) Increasing doses of transfected *AS-Uchl1* titrate UCHL1 protein but not mRNA levels in HEK cells. The increase in UCHL1 protein levels in chart (b) and (c) could be observed in the gel images shown in the boxes, where beta-Actin is used as control. Also the data shown in (b) and (c) indicate mean ± s.d., n ≥ 3. *(Above set of figures are taken from Carrieri et al., 2012).*

Subsequently, synthetic antisense lncRNA with identical domain composition as that of *AS-Uchl1*, referred as synthetic SINEUPs (Zucchelli et al., 2015a) were also shown to act similar to that of natural *AS-Uchl1*. This suggests for a possible functional relationship between SINE repeats and antisense lncRNA. Interestingly, when the activity of synthetic SINEUPs were tested across the cell lines of multiple species including mouse, Chinese hamster, human and monkey (Patrucco et al., 2015; Yao et al., 2015; Zucchelli et al., 2015a), they showed a similar increase in translation of endogenous cellular mRNAs. The successful design of synthetic SINEUPs and their testing across multiple species has opened up many new possibilities for their applications, such as, in protein manufacturing, usage as reagents in molecular biology experiments and most importantly in RNA therapeutics discussed in detail in a recent review (Zucchelli et al., 2015b).

Currently, antisense lncRNAs are emerging as an important class of lncRNAs that are being extensively studied and scrutinized for various regulatory activities. Given that they share an overlapping region with respect to sense coding gene, the existence of a possible mechanism of mutual regulation is highly expected. In sum, the post-transcriptional protein up-regulatory activity of the antisense lncRNA *(AS-Uchl1)* explained by *Carrieri et al.*, adds a new layer of complexity to the existing molecular network of activities that involves the embedded sequence of SINE repeat as an important functional domain *(also referred as effector domain by Zucchelli et al., 2015a)*. Currently, the underlying molecular mechanism for the translational up regulatory activity of inverted SINE in *AS-Uchl1* is not known, hence it is open for further research and experimentation.

1.6. Transposable elements

We just discussed about the functional association of SINE repeat which is a transposable element (TE), with that of *AS-Uchl1* lncRNA, in the translation regulation of overlapping sense Uchl1 gene. Interestingly, TEs themselves are known to be involved in several gene regulatory activities and to be intimately linked to eukaryotic genomes for millions of years playing key role in the evolution of the genomes (well reviewed in Bourque, 2009). Could the reported activity of *AS-Uchl1* lncRNA which requires a SINE TE as an essential functional domain, be a new addition to the existing list functionalities for all TEs? To know this, it is important to introduce TEs in detail and get familiarized with their different classes, mode of action and their impact on gene regulation and evolution.

1.6.1. Introduction to TEs

TEs were first discovered by Barbara McClintock in 1940's during her extensive genetic analyses on maize. She was particularly interested on spotted corn kernels, and asked a simple question, "why the spotted corn kernels were spotted?". She basically wanted to understand the genetic mechanism responsible for the phenotype of spotted corn kernels. McClintock discovered the spotted corn kernels were caused by a new type of genetic elements addressed as *transposable elements (TEs)*, that could give rise to reversible mutations in the genes involved in the pigment biosynthesis process. Briefly, McClintock identified TEs are the mobile genetic elements that gets inserted into the gene involved in pigment biosynthesis thereby disrupting its function resulting in the formation of un-pigmented areas in the corn kernels. But in later developmental stages of kernels, TEs can get excised off from the gene, thereby restoring its expression and functionality, resulting in the formation of the pigmented

areas in the corn kernels giving it a spotted appearance. Based on these interesting observations, McClintock proposed two main functions for TEs - 1) insertional mutagens and 2) "controlling elements" that can regulate the expression of nearby genes (Mcclintock, 1951). At the time her claims were not universally accepted by the scientific community, as a result TEs remained under the cover of "Junk" or "Selfish" DNA until 1990s. However, during this time a large body of knowledge were accumulated. For example, *Britten & Kohne,* in 1968, confirmed the presence of greater percentage of repeated sequences in higher organisms using the reassociation *of DNA* method. Subsequently, *Grimaldi & Singer,* in 1982 identified an *Alu* sequence in monkey which at the time, was already known as the dominant Short Interspersed Elements (SINEs) in primates. This identified *Alu* SINE TE was flanked by 13 bp duplication of known sequence of a satellite suggesting a possible mechanism for mobility. With the subsequent pileup of similar evidence, Barbara McClintock's work eventually got recognized and she was awarded with the Nobel Prize in Physiology and Medicine for the discovery of TEs in 1983.

Later, *Kazazian et al., 1988* demonstrated for the first time that a TE insertion in human genome could cause disease. *Batzer et al. 1996,* made use of polymorphic *Alu* insertions as a unique source of nuclear genetic variability for the investigation of human population genetics and surveyed of 14 human population group across continents. The results of this survey indicated that the genetic variation between European population were smaller than the genetic variation observed between Africans (Batzer et al., 1996). Further, *Moran et al.* In 1999, showed evidence for the retrotransposition of L1 elements, which is an Long Interspersed Element (LINE), could act as a vehicle to mobilize non-L1 sequences such as exons or promoters from their 3' flanks to new genomic locations leading to exon shuffling (Moran,

DeBerardinis, & Kazazian, 1999) *(details on Alu SINE and L1 LINE TEs mentioned in this section and the position they hold in the standard TEs classification system is discussed in next section).* In followup to these initial discoveries of TEs and strong evidence of their mobility and influence to the structure of their host genomes, a large number of similar exciting discoveries were made, where TEs were found to be involved in disease, epigenetics, gene regulations and evolution (well reviewed in Reilly et al., 2013; Biemont, 2010; Feschotte, 2008)

1.6.2. Classification of Transposable Elements and characteristics of their transposition

TEs are the genomic sequences capable of moving from one location in the genome to another through a process called transposition. They can be broadly classified into two major classes based on their mechanism of transposition - 1) The DNA transposons (also known as class II TEs) which move via cut-and-paste mechanism and 2) RNA transposons (also called as retrotransposons, or class I TEs) which move via copy-and-paste mechanism using RNA as an intermediate. The retrotransposons can further be classified into two sub-types I) LTR (Long Terminal Repeats) such as ERVs (endogenous retrovirus) and 2) the non-LTR retrotransposons such as LINEs (Long Interspersed Elements) and SINEs (Short Interspersed Elements). The L1 elements are the only known active and autonomous LINE elements in the human genome (reviewed in Beck et al., 2013), whereas *Alus* in human and B1, B2 and B4 elements in mouse and MIRs together accounts for different known SINE elements.

The retrotransposons can also be broadly classified as autonomous and non-autonomous, based on their ability to mobilize. Autonomous retrotransposons are LTRs and LINEs which can code for the proteins and enzymes that are required to mediate their transposition and

integration into the new genomic sites. In contrary, non-autonomous retrotransposons such as SINEs, lack open reading frames to code for such proteins and usually rely on the protein and enzyme machinery produced by autonomous retrotranspsons for their transposition and integration into the new genomic sites. To understand better how different TEs are transposed and regulated in the genome, it is also necessary to analyze and understand the polymorphism in their nucleotide sequences (Lerat et al., 2003; Lerat et al., 1999), because TEs tend to either degenerate by truncations, deletions, insertions, substitutions and can eventually eliminate from the genome, or such events could lead to the divergence of TEs into multiple TEs families. It is important to note that the events of truncation, deletion and insertion of other elements could result into the nests of elements or partial TEs (TEs with incomplete sequence) which are most often abundantly identified across the eukaryotic genomes (Kaminker et al., 2002; Lerat et al., 2003).

1.6.3. Transposable Elements in gene regulation

TEs are known to be involved in gene regulation in several ways, for example they can either act as enhancers or may serve as the alternative promoters or even as the mobile sites for transcription factor binding to create novelty in the transcriptional regulatory networks. One of the established examples of TEs acting as enhancers include the work of *Bejerano et al. 2006,* who in 2006, identified a class of conserved, primarily non-coding regions in tetrapods that originated from a previously unknown SINE retroposon family which was active in the Sarcopterygii (lobe-finned fishes and terrestrial vertebrates) in the Silurian period at least 410 million years ago. Using a mouse enhancer assay they showed that the non-coding sequences originated from a SINE element, acted as a distal enhancer to a neuro-developmental gene *ISL1* from a distance of 0.5 million bases (Bejerano et al., 2006). Similarly, *Sasaki et al* in

2008, demonstrated that a particular TE from yet another SINE family conserved among the genomes of Amniota (mammals, birds, and reptiles), present in a distance of 178 kbp from FGF8 (fibroblast growth factor 8) gene, acts as enhancer and recapitulates FGF8 expression in two regions of the developing forebrain, namely the diencephalon and the hypothalamus (Sasaki et al., 2008). These examples together demonstrates the role of SINEs as enhancers in the regulation of genes that are involved in the development of mammalian neuronal network.

As a part of FANTOM4 project, in 2009 *Faulkner et al.*, showed that between 6 and 30% of the cap-selected mouse and human transcripts were initiated from within repetitive elements and that retrotranposons present immediately to the 5' ends of the protein-coding loci acted as alternative promoters. They also identified more than a quarter of the RefSeq transcripts possessing a retrotransposon at their 3' ends were associated to a strong evidence for reduced expression compared to rest of the transcriptome. Additionally, with the genome wide screening they identified 23,000 candidate regulatory regions that were derived from retrotransposons (Faulkner et al., 2009). Another study by *Pereira et al* in 2009, which aimed at identifying if the new TE insertions could affect gene expression divergence between mouse and rat, showed a strong correlation in expression divergence and differential presence of retrotransposons such as SINEs and LTRs. At the same time they also identified no significant correlation in case of ancestral LINE retrotransposon (Pereira, Enard, & Eyre-Walker, 2009) suggesting for differential TE dynamics. Indeed, there are several studies which indicates that each TE class is different from one another and likely play different functional roles in transcriptional regulation. For example, a study carried out on human genome by *Thornburg et al. 2006,* showed that SINE elements which are GC-rich, are predominantly present near promoter and genic regions, whereas LINEs which are AT-rich and are found

usually in gene poor regions of the genome (International Human Genome Sequencing Consortium, 2001; Thornburg, Gotea, & Makałowski, 2006). LINE elements, particularly L1 (LINE-1) are known to contain YY1 (Yin Yang 1) protein binding site which is proposed to function as a component of the LINE-1 core promoter to direct accurate transcription initiation. (Athanikar, Badge, & Moran, 2004), RNA polymerase II (pol II) promoter to direct accurate initiation of transcription by the RNA polymerase II machinery and antisense promoters that have been shown to influence the transcription of adjacent genes (Speek, 2001). The L1 elements are also known to contain 14 over-represented classes of transcription factor binding sites, double the number of transcriptional signals in SINE elements. Unlike L1, SINE elements do not contain pol II promoters, instead the active SINE elements are transcribed by RNA pol III and carry pol III promoter (Kramerov & Vassetzky, 2011; Varshney et al., 2015), that do not influence the transcription of protein coding genes. Among all TE classes, LTRs are known to carry the binding sites for almost all transcription factor classes. Unlike SINEs, the number and proportion of LTR and DNA TEs in promoter regions is the lowest, probably due to their higher divergence and fragmentation which makes their detection harder or even impossible, with similarity searching techniques. Taken together, TEs have a potential to influence gene regulation at genomic scale by carrying transcription regulating signals. When they are inserted in promoter regions, they can alter gene expression patterns by contributing transcription factor binding sites that are previously not present in promoters of specific genes. (Thornburg et al., 2006).

Some of the most interesting and recent evidence for TEs influencing gene regulation include, the first genome wide study of gene expression variation due to TEs in *Drosophila,* which revealed that the TE insertions in or near the transcript is significantly associated with

reductions in its expression levels (Cridland, Thornton, & Long, 2015). Studies have also revealed that the human embryonic tissues expresses the greatest diversity of TE-associated TSSs, highlighting the potential of TEs to drive cell type and developmental stage-specific gene expression, particularly during early embryogenesis when the genome becomes demethylated (Messerschmidt, Knowles, & Solter, 2014).

1.6.4. Transposable Elements in genomic rearrangements and genome evolution

Apart from influencing the transcriptional regulatory networks, TEs are also involved in the evolution by facilitating chromosomal rearrangements such as deletions, insertions, duplications, inversions and recombination of the host genomes. These aspect of TE functions also has implication for understanding several human genomic disorders (reviewed in Lupski, 1998). The two well established mechanisms by which TEs associated chromosomal rearrangements can occur are - 1) homologous recombination or 2) alternative transposition process. TEs are involved indirectly in homologous recombination process, where they present the genome with multiple similar sequences between which the recombination can occur or by faulty repair of double-strand breaks formed during transposable element excision using ectopic homologous sequences as a repair template (Gray, 2000). On the other hand, two closely-located TEs can induce chromosomal breakage and rearrangements via alternative transposition mechanism, where TEs induce chromosomal rearrangements directly by an alternative version of the traditional transposition reaction. For example, in case of class II TEs, the first step of transposition is the synapsis of complementary left and right ends followed by the excision of the ends which result into the target site capture and strand transfer, but in case of alternative transposition mechanism, the complementary ends from

separate TEs synapse rather than the traditional synapsis of complementary ends from a single TE (Zhang et al., 2014; Gray, 2000).

There are several line of studies that explain the role of TEs in the chromosomal rearrangements and structural variation of the genomes, for example, in human genome *Alu* elements are found to promote homologous recombination events as they provide short regions of homology at adequately frequent intervals. And the unequal crossing over between *Alu* elements on homologous chromosomes can result in heritable duplications and deletions of the intervening regions (Prak & Kazazian Jr, 2000). *Alus* are also known for the retrotransposition-mediated deletion (ARD), by endonuclease dependent or independent mechanisms, of a portion of adjacent sequence occasionally larger than the *Alu* insert itself (Callinan et al., 2005). However, the initial evidence for a TEs retrotransposition-mediated deletion was derived from a study by *Gilbert et al in 2002*, who demonstrated using the cultured human cells that the L1 retrotransposition event can generate large target site deletions (Gilbert, Lutz-Prigge, & Moran, 2002). Further, an analysis of primate genomes *(human, chimpanzee and rhesus macaque)* for endonuclease-independent *Alu* insertions, suggests that the *Alu* insertions are involved in DNA double-strand break repair (Srikanta et al., 2009). Additionally, *Alu* insertions into coding or regulatory regions are also known to alter the architecture of a gene which might be deleterious depending upon the insertion location and the affected gene (Deininger, 2006). For example, several line of studies focusing on SINE *Alus* revealed the evidence for their exonization and involvement in alternative splicing (Gal-Mark, Schwartz, & Ast, 2008; Shen et al., 2011; Sorek, Ast, & Graur, 2002).

TEs are also known to mediate segmental duplications in the genome. A comparative genome study by *Bailey et al* in 2003, showed human genome is enriched for large homologous segmental duplications. A systematically examination of the sequence features at the junctions of the duplications revealed that the *Alu* short interspersed elements were significantly enriched near or within these duplicated junction (Bailey, Liu, & Eichler, 2003). Similarly, in another comparative analysis between mouse and human genomes revealed that, in contrast to SINE enrichment at the boundaries of segmental duplications in human, mouse segmental duplications are enriched for LINE and LTR TEs (She .X et al., 2008). This example also suggests that the TEs have adapted lineage-specific functionalities for the benefit its host genome. TEs indeed are the most ineage-specific elements in eyukaryotic genomes, for example, a comparative study between mouse and human revealed that the regulatory sequences that are contributed by TEs are exceptionally lineage specific where the majority of TE-derived *cis*-regulatory sequences in the human genome come from the relatively younger *Alu* and L1 TE families. However, none of these TE derived *cis*-regulatory sequences are conserved between the human and mouse genome. This suggests, in human the TE insertions generated regulatory sequences that occurred after the human and mouse evolutionary lineages diverged (Mariño-Ramírez et al., 2005).

1.6.5. Transposable Element regulation

Until now, we discussed very briefly about various ways through which TEs take control of gene regulation and how TEs reshape their host genomes leading to the evolutionary changes. Given that, TEs are the powerful facilitators of the genome evolution, their uncontrolled activation could result into genomic instability that potentially could make their host vulnerable. As a consequence, eukaryotic genomes have evolved sophisticated mechanisms to

check TEs activities. Before we discuss about these mechanisms, it is important to know that TEs are active in both germline and somatic cells. For a long time TEs were thought to have a restricted activity to the germ cells or early embryonic cells, however the most recent reports indicate that somatic L1 retrotransposition could also occur later in development. For example, over the past decade, several studies have shown that the mammalian brain, particularly cells of the neuronal lineage can express L1 RNA and that the somatic retrotransposition of engineered L1 elements can take place in transgenic mouse models (An et al., 2009; Kano et al., 2009). Another study focusing on the identification of insertional sites of L1, *Alu* and SVA elements in the brains of three individuals revealed, 7,743 putative somatic L1 insertions in the hippocampus and caudate nucleus along with the 13,692 somatic *Alu* insertions and 1,350 SVA insertions. The results of this study also demonstrated that the retrotransposons mobilize to protein-coding genes that were deferentially expressed and active in the brain. These observations suggests for somatic genome mosaicism driven by retrotransposition that may have reshaped the genetic circuitry responsible for normal or abnormal neurobiological processes (Baillie et al., 2011). Other recent studies striving to identify and understand the L1 somatic insertions maps are discussed in a recent review by Richardson, Morell, & Faulkner, 2014.

In sum, somatic TE retrotransposition is an event which is recently identified and currently is a hot topic of research. In contrary, the germline TE retrotranpositions are well studied, hence there are also well established mechanisms that are known to limit TE activities in germ cells. One of the well known mechanisms for the regulation of TEs in germline of eukaryotes is DNA methylation, which is an essential epigenetic modification largely restricted to the CpG dinculeotides and serves as a repressive mark on the gene expression patterns (well reviewed

in Jaenisch & Bird, 2003). In somatic tissues, the CpG methylation landscape is relatively static that exhibits global patterns based on relative CpG density. The CpGs at the promoters of the housekeeping or developmental genes are largely unmethylated, whereas the CpGs at non-regulatory in the genome are largely methylated and only a small fraction of CpGs switch their methylation status as a part of gene regulatory event (Suzuki & Bird, 2008; Meissner et al., 2008). On the other hand, studies have suggested that the germline undergoes two genome-wide DNA demethylation events, the first is immediately after the fertilization of the zygote and the second during the establishment of the primordial germ cells, which are the direct progenitors of sperm and oocytes (reviewed in Seisenberger et al., 2013; Smith et al., 2012). When the DNA is methylated by the transfer of a methyl group to the 5^{th} carbon of cytosine, the LTR and non-LTR TEs are silenced due to the transcription suppression of retrotransposon RNA, whereas during demethylation phases TE-associated TSSs are expressed to drive cell type and developmental stage-specific gene expression. This demonstrates that the DNA methylation mechanism controls TE activities in a manner beneficial for the host genome and development of the embryo (reviewed in O'Donnell & Burns, 2010; Messerschmidt, Knowles, & Solter, 2014). Apart from silencing of transposable elements, DNA methylation is also involved in the regulation of gene expression, genomic imprinting, and X-chromosome inactivation (Li & Zhang, 2014). Similarly, other mechanisms which controls TE activities in germline include piRNA biogenesis shown in *Drosophila*, where the piRNAs associate with piwi proteins serve as guides that lead to the cleavage of expressed transposon targets (O'Donnell & Boeke, 2014), thereby immunize the *Drosophila* germline against potentially sterilizing transposition events.

1.6.6. Abundance of TEs in lncRNAs

So far we discussed regarding different classes and characteristics of TEs, their several mode of action in gene regulation and chromosomal rearrangements. We also discussed briefly about their role in lineage specific genome evolution and how TE actives are highly controlled in the genomes to avoid genomic insatiability due to unchecked TEs insertions. Altogether, this already demonstrates that the TEs are the major contributors to the evolution of genomes. However, another important aspect of TEs which still needs to be addressed is their abundance in eukaryotic transcriptomes. TEs are already well known to occupy a large fraction of many eukaryotic genomes (Feschotte & Pritham, 2007). Interestingly, the large scale identification, characterization and functional elucidation of lncRNAs have clearly shown that the TEs constitute a major portion of lncRNAs in the eukaryotic transcriptomes, in contrast to protein coding genes. For example, a study by *Kelley & Rinn, 2012* identified the TE composition of human and mouse lncRNAs by intersecting the TE annotations with a catalog of 9,241 human and 981 mouse lncRNAs and identified, although TEs comprise less lncRNA sequence with respect to genomic background, they contribute substantially more to the lncRNAs sequences than protein coding genes in human *(Figure 1.3 a)*. Similar observation was also seen in case of mouse, where TEs are depleted overall among lncRNAs relative to the genomic background frequency but a substantial 33% of lncRNA sequence are TE-derived *(Figure 1.3 c)* and this percentage is much higher with respect to protein coding sequences. In addition *Kelley & Rinn, 2012,* also identified a non-random distribution of many TE families which are significantly enriched among lncRNAs with respect to genomic background sequences, for example, L1 elements are found to be significantly depleted whereas ERV1s are enriched among lncRNAs promoters when compared with genomic background sequences in both human *(Figure 1.3 b)* and mouse *(Figure 1.3 d)*.

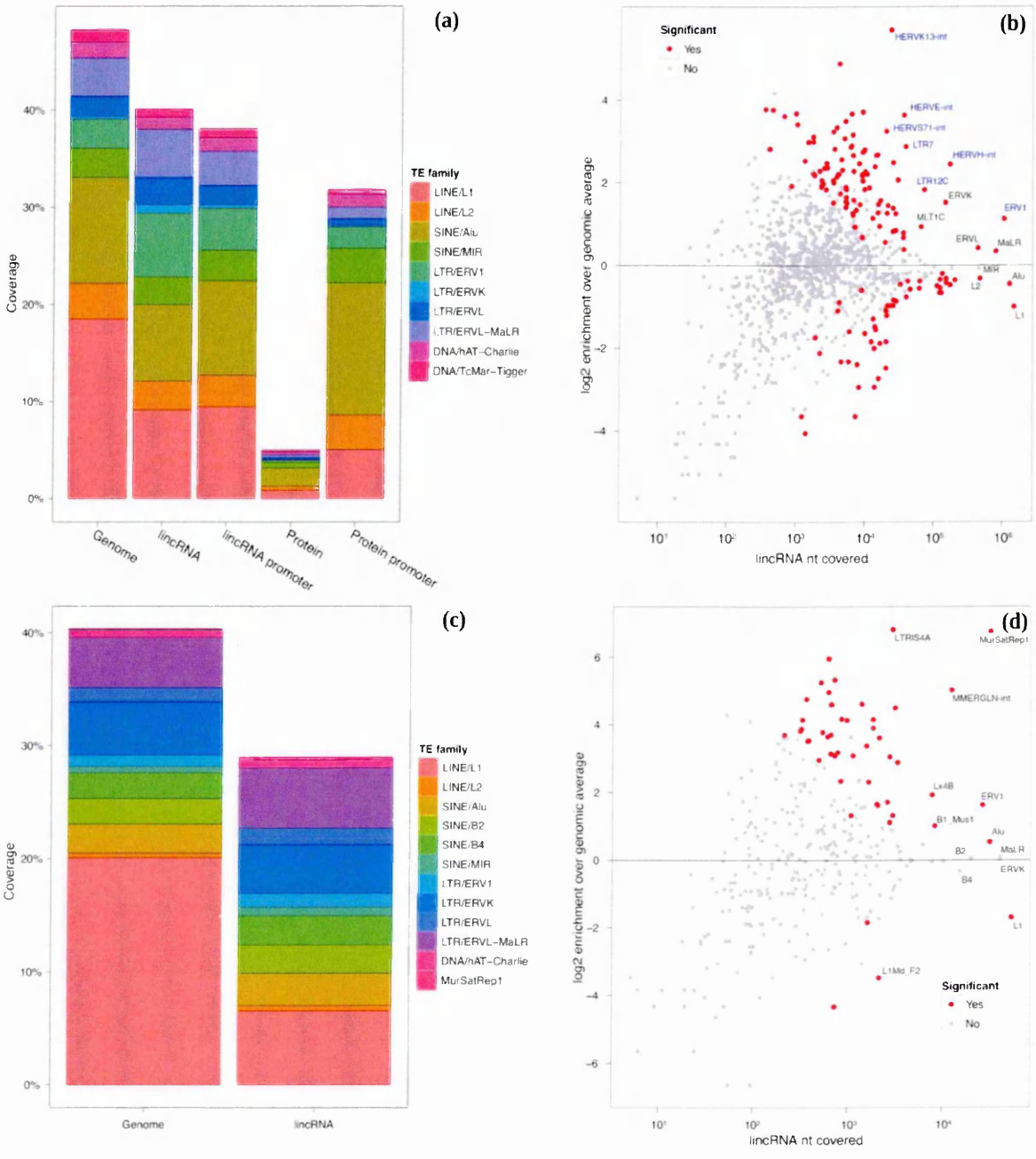


Figure 1.3 | TEs composition of lncRNAs. in human and mouse. Charts (a) and (c) represents the sequence level TE coverage percentages among human and mouse respectively. In case of human the coverage is shown for genome, lncRNAs, protein-coding genes and their promoters whereas for mouse TE coverage is shown

across the genomic and lncRNA sequences. Charts (b) and (d) represents the coverage enrichment of specific TE families in human and mouse respectively. Enrichments are above zero on the y-axis, and depletions are below zero. The significantly enriched TEs are shown in red dots. *(Above set of figures are take from, Kelley & Rinn, 2012).*

Similarly, other studies that focused on the lncRNA sequences catalog collected from GENCODE (Derrien et al., 2012), Cabili et al., 2011, Pauli et al., 2012 and Ulitsky et al., 2012, for human, mouse and zebrafish respectively, have shown that a substantial fraction of the lncRNAs contain exonized TE sequences. And specific TE families such as LTR/ERV in human and mouse and DNA transposons in zebrafish are over-represented among lncRNAs (Kapusta et al., 2013). Together these studies suggests for the existence of a potential functional association between the lncRNAs and TEs.

Based on our discussions on TEs and lncRNAs it is noted that they both form a complex but interesting layer of eukaryotic transcriptomes. We have seen, both have their own independent regulatory effects on the transcriptome activities. However, we have also recently witnessed their coordinated effect on the translation regulation of proteins at post-transcriptional level (Carrieri et al., 2012; Zucchelli et al., 2015a), which suggests a new level of molecular complexity that yet requires further explorations. The TEs embedded regions in the lncRNAs have also been proposed as key element in the RIDL hypothesis (Repeat Insertion Domains of LncRNAs) which propose that exonized TEs might constitute functional domains of lncRNAs (Johnson & Guigó, 2014). A growing number of RIDLs have been experimentally identified whereby the TE derived sequence of lncRNAs act as RNA-, DNA-, and protein-binding

domains/motifs. Based on this, it has been hypothesized that the inserted TE sequences are repurposed as recognition sites for both protein and nucleic acids, reflecting a more general phenomenon of exaptation during lncRNAs evolution. The RIDL hypothesis also has the potential to explain how functional evolution can keep pace with the rapid gene evolution observed in lncRNA (Johnson & Guigó, 2014), which is important to understand, how can lncRNA genes which are born over relatively short evolutionary timescales, rapidly acquire molecular activity and play new functional roles? Several studies have addressed the evolution of lncRNA genes and their functional activities (Ponting, Oliver, & Reik, 2009), however the processes governing their functional evolution have not often investigated in detail. TE are likely to have contributed to both processes (Johnson & Guigó, 2014). The identification of functional TEs domain in lncRNAs or in other words the candidate RIDL elements would rely on the criteria like

- *Base-level over-representation* – Determining the TEs over/under representation within the lncRNA exons might reflect the effect of TEs functionality, as the potent TE fragments may be selected against in many lncRNA hosts where their presence are inappropriate or detrimental to function and only maintained in only a subset of lncRNAs where they contribute a selective advantage
- *TE subregion overrepresentation in lncRNA* - TEs tend to insert only a sub-fragment of their consensus motif during an event of a novel insertion that often have variable lengths originating at the 3′ end due to incomplete reverse transcription. Hence, it is expected that they will be over-represent in the exons of host lncRNAs with respect to a subset of other lncRNAs or the genome as a whole. Identification of such over-represented TE sub-regions might be useful method to filter functional lncRNA domains originating from TEs, where TEs might hold clues to their role in lncRNAs

- *Strand bias* - If the function of a TE depends on the strandedness in which it is transcribed, then the TEs might preferentially tend retain their particular strand orientation relative to the host lncRNA exon. Hence investigating for the strand bias of TEs within lncRNAs might help to filter out a subset of lncRNAs that have their functional domains originating particularly from the TEs strandness. A crucial consideration in these cases is that a strong strand bias will be expected, where the TEs are contributing to the structures of lncRNAs (Kapusta et al. 2013)
- *Secondary structures* – Many TEs may contain secondary structures that mediate their activity and hence looking for their over-representation of of structured sequence might hold clues to their functional role in a subset of lncRNAs
- *Cellular localization* - TE RNAs in isolation tend to localize at different sites within the cell (Goodier et al. 2010) hence it can be expected that the the signal driving this localization presumably would act on lncRNA hosting such same TEs. Hence investigating for the differential cellular localization of TEs might also help to separate lncRNA subsets that localize to specific locations where they play their functional roles (Johnson & Guigó, 2014).

1.7. Aims and synopsis of my PhD project

Inspired by the findings of *Kelley & Rinn, 2012* and *Carrieri et al., 2012*, my PhD thesis aims to gain broader insights on antisense lncRNAs and to define functional association between the antisense lncRNAs and TEs among vertebrate and invertebrate species.

Based on the described modular nature of the *AS-Uchl1* by *Carrieri et al., 2012 (Figure 1.2)*, and the need to identify if such a characteristic could be widespread across multiple antisense lncRNAs *(referred as ASlncRNA from here on)* in the transcriptomes of multiple species, I established the following objectives -

1. To identify all existing ASlncRNAs in the transcriptome and to determine if they are biased towards the content of SINE elements in comparison to other TEs, because the effector domain of *AS-Uchl1* is an inverted SINE element
2. To analyze the modular nature of the identified ASlncRNAs by scrutinizing their characteristics of the 5' binding domain and the 3' effector domain and explore their functional implications over the sense overlapping coding genes

During the course of my study I have developed a profound understanding of various computational aspects, tools and the usage of large datasets from the public data repositories. For the precise identification of the overlapping features in the transcriptome, characterization of ASlncRNAs, mapping repeat elements to the transcripts and to compute the TE coverage percentages among the classified transcript groups, I have developed a complex and flexible bioinformatic pipeline.

Using this pipeline, I have analyzed the transcriptomes of three vertebrates – 1. Human *(Homo sapiens)*, 2. Mouse (*Mus musculus*), 3. Zebrafish (*Danio rerio)*, and two invertebrates – 4. Fruit-fly *(D. melanogaster)* and 5. Worm (*C. elegans)*, and have identified a large number of ASlncRNAs particularly in human and mouse. To analyze the enrichment of TEs among ASlncRNAs, for the graphical representations and for functional enrichment analysis, I have developed different modules of the pipeline which can perform randomization and comparative analysis by applying statistical tests to compute the significance of the observed TEs coverage and functional annotations.

To scrutinize the characteristic of the 5' binding domain such as the effect of the presence or absence of ATG (translation initiation site, TIS) within the overlapping region of the binding domain over the functional association of sense coding genes, I performed ATG overlap based grouping of S/AS pair of coding genes and analyzed their functional enrichment. Further, using these group of genes, I analyzed the possible changes in the cellular localization of their corresponding peptides, considering the full-length and the truncated peptide sequence which would correspond to the translation of the mRNA from a secondary TIS, downstream and outside of the ASlncRNA overlap region.

To analyze the characteristic of the 3' effector domain, such as, the effect of SINE orientations with respect to the ASlncRNA over the sense coding genes, I performed the classification and grouping of mRNA based on the SINE orientations in ASlncRNA partners, followed by functional enrichment analysis. Further to understand how the coding genes and the overlapping ASlncRNA carrying SINE repeats with specific orientation would react under cellular stress conditions, I made use of a recently published expression data collected from

polysome profiling and fractionation experiment on the human MRC-5 cell-lines in stress and normal conditions and performed an overlap analysis. Altogether, using the available resources and high valued data from public data repositories, I have generated the results which shed light on the previously unanalyzed areas of ASlncRNAs. Additionally, I have produced a modular pipeline which can be deployed for analyzing the ASlncRNA and their repeat content across the transcriptomes of multiple species which have a well annotated set of lncRNAs.

Chapter 2

A bioinformatic pipeline for the identification of ASlncRNA and computation of their TEs coverage

2.1. Introduction

Previously, we discussed about the functional activity of *AS-Uchl1* in upregulation of UCHL1 proteins at post-transcriptional level during cellular stress conditions in mouse MN9D dopaminergic cell lines, where its activity mainly relies upon its modular nature *(Figure 1.2).* The search for similar such natural ASlncRNAs revealed, *AS-Uxt* antisense to the ubiquitously expressed transcript *(Uxt)* gene contain a similar SINEB2 element near to its 3' end, that also shows an identical regulation of UXT protein level without affecting *Uxt* mRNA expression in mouse MN9D cells. This suggests, for the existence of multiple natural modular ASlncRNAs similar to *AS-Uchl1* awaiting to be discovered. Interestingly, the artificial constructs of *AS-Uchl1* (synthetic SINEUPs) targeting specific genes by swapping only 5' overlapping complimentary sequence according to the target mRNA and retaining 3' domain intact that contains the inverted SINEB2 also behaved identical to the full-length *AS-Uchl1*. This indicates that the 3' domain containing inverted SINEB2 element is an important effector domain for the protein upregulation activity, whereas the 5' overlapping region of the ASlncRNA is necessary to provide the specificity to target coding genes (Carrieri et al., 2012; Zucchelli et al., 2015a).

To determine if the observed activity of *AS-Uchl1*, *AS-Uxt* and synthetic SINEUPs could be a widespread phenomenon for all similar naturally occurring ASlncRNAs in eukaryotes, I performed a transcriptome wide identification, characterization and functional inspection of ASlncRNAs. The transcriptome wide identification of ASlncRNA for multiple eukaryotic species seeks an automated bioinformatic pipeline capable of determining the overlapping genomic features (transcripts) from a given genome and its transcriptome. Currently, there are multiple tools available for this purpose, such as *Galaxy*, a feature-rich web interface (Giardine, Riemer, & Hardison, 2005; Goecks, Nekrutenko, & Taylor, 2010), *BEDTools*, a UNIX command-line interface (Quinlan & Hall, 2010), *BEDOPS*, a computational memory efficient command line interface (Neph et al., 2012) and *IRanges*, an R programming platform based Bioconductor package. *IRanges* is particularly very useful for computing on annotated genomic ranges and integrating genomic data, along with the integration of *GenomicRanges* and *GenomicFeatures* Bioconductor libraries (Lawrence et al., 2013).

I chose to use the R programming language which is a software environment for data manipulation, statistical computing and graphics, to developed a bioinformatic pipeline which could be used to identify and study transcriptome wide ASlncRNAs. R also integrates Bioconductor infrastructure that provides tools for the analysis and comprehension of high-throughput genomic data. The Bioconductor infrastructures extensively used for genomic feature operations in my pipeline are compiled with the three core libraries- 1) IRanges, that provides the fundamental range data structures and operations, while 2) GenomicRanges, builds upon IRanges to add biological semantics to the metadata, including explicit treatment of sequence name and strand and finally 3) GenomicFeatures, that enables the access to the gene models and other annotations (Lawrence et al., 2013). The close integration of other R

packages, extensions, in-memory data structures, vast graphic features for the data representation, stable interaction with external tools and an active support from the Bioconductor community makes R a suitable environment for multiple in-line analysis and for the development of bioinformatic pipelines (Gentleman R. C., et al., 2004; Gentleman R., 2009; Ross Ihaka & Robert Gentleman, 1996).

In this chapter, I aim to describe in detail the workflow of the bioinformatic pipeline I have developed to identify, characterize, and analyze the ASlncRNAs in the transcriptome of multiple species including vertebrates and invertebrates. I have also described the initial set of results of the comparative TEs coverage enrichment analysis, that I performed by implementing the modular pipeline. This analysis mainly aimed to determine if the TEs, particularly SINE elements are the integral components of ASlncRNAs as seen in the case of *AS-Uchl1*, in contrast to rest of the lncRNAs,.

2.2. Material and Methods

2.2.1. Structure of the pipeline

The modular pipeline for the identification of ASlncRNAs is implemented on a collection of R scripts *(R version v3.2.2 as per 2015-08-14)*, taking advantage of several Bioconductor libraries https://www.bioconductor.org/, and CRAN (Comprehensive R Archive Network) extensions https://cran.r-project.org/. Some of the core implemented libraries, include *IRanges (v2.0.1)*, for genomic feature operations (Lawrence et al., 2013), *biomaRt (v2.22.0)* for data retrieval from Ensembl (Durinck et al., 2005), *data.table (v1.9.6)* (Dowle, Short, Lianoglou, & Srinivasan, 2014) and *reshape2 (v1.4.1)* (Wickham, 2014) for data handling and processing,

ggplot2 (v1.0.1), for data representation (Wickham, 2009), *Parallel*, for the parallel computation (Ripley, Tierney, & Urbanek, 2015).

The work-flow of the pipeline can be broadly divided into three different phases *(Figure 2.1)*

1. Data collection (Downloading all the required data)
1. Data processing (S/AS Identification, repeats mapping and characterization of ASlncRNAs)
2. Data analysis and representation

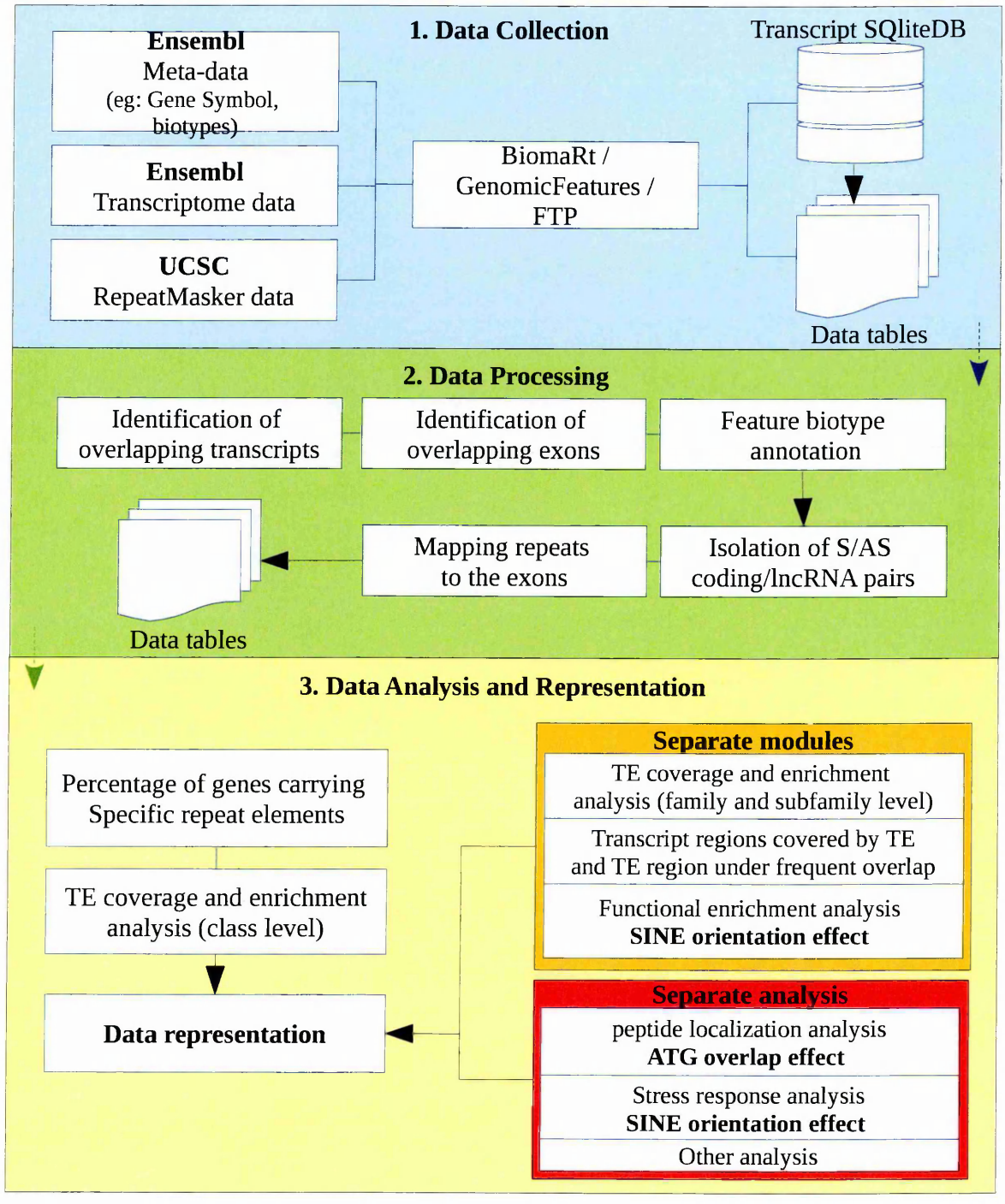


Figure 2.1 | Pipeline work-flow The above figure represents the work-flow and the modular structure of the pipeline.

2.2.2. Data collection

Given that, SINEB2 element is the effector domain essential for up-regulating the protein levels of the coding genes overlapping to *AS-Uchl1*, *AS-Uxt* and synthetic SINEUPs, I was interested to unveil if similar ASlncRNAs are generally biased towards SINE repeat coverage with respect to rest of the transcriptome in eukaryotics. For this, I selected to analyze the transcriptomes of three vertebrates – 1. Human *(Homo sapiens)*, 2. Mouse (*Mus musculus*), 3. Zebrafish (*Danio rerio)*, and two invertebrates – 4. Fruit-fly *(D. melanogaster)* and 5. Worm (*C. elegans)*, as they are the extensively studied species in the projects undertaken by Encode, mouse Encode and modEncode, and also have a well annotated catalog lncRNA genes *(Table 2.1)*.

The transcriptome data from these species are primary requisite for any analysis in the pipeline. Transcript mappings on the genome and relative annotations were downloaded using biomaRt, corresponding to the specific Ensembl release versions *(Table 2.1)* and stored locally into an SQLite database addressed as the "TranscriptDB", constructed using the parser available in the *GenomicFeatures* library. The constructed database contain the information regarding transcripts such as the chromosome number, start and end coordinates, strand, number and location of exons, exon ranks, CDS (coding DNA sequence) start, end coordinates, which are all properly managed by the *GenomicRanges* infrastructure (Carlson, Aboyoun, & Pages, 2015). Other essential information corresponding to transcripts, for example, gene symbols, homolog genes, gene ontology annotations and most importantly, gene and transcript biotypes *(corresponding to GENCODE in case of human and mouse)* are also downloaded from the Ensembl database via biomaRt and systematically organized into data tables for later references. For mapping repeat elements to the transcripts, I used

RepeatMasker annotations downloaded via FTP from the UCSC genome browser (ftp://hgdownload.cse.ucsc.edu/goldenPath) *(Table 2.1)*. The data download and organization steps are automatically performed by the data collection module of the pipeline that requires only the species names as input (*Figure 2.1).*

Dataset	*Human* *Homo sapiens*	*Mouse* *Mus musculus*	*Zebrafish* *Danio rerio*	*Fruit Fly* *D. melanogaster*	*Worm* *C. elegans*
Transcriptome	Ensembl 82	Ensembl 82	Ensembl 82	Ensembl 82	Ensembl 82
Genome Assembly	GRCh38.p3	GRCm38.p4	GRCz10	BDGP6	WBcel235
UCSC (RepeatMasker)	hg38	mm10	danRer10	dm6	ce10
Gencode	Gencode 23	Gencode M7			
Genes					
Total coding genes	22017	22158	25465	13918	20477
lincRNA	7958	3515	855	2366	169 + 7687 (ncRNA)
antisense	5722	2146	671		
Processed transcript	800	750	1142		
Total lncRNA genes	14480	6411	2668	2366	7856
Transcripts					
Total coding transcripts	87256	50558	31389	30353	30939
lincRNA	14571	5272	801	2776	176 + 8054 (ncRNA)
antisense	11490	3133	645		
Processed transcript	29867	13500	3116		
Total lncRNA transcripts	70408	21905	4582	2766	8230

Table 2.1 | Dataset used in the study. The above table contains the details of datasets used in the study, including total number of coding and lncRNA genes (yellow) and transcripts (green) present in the transcriptomes of each species analyzed using the pipeline. The total lncRNAs are represented by the sum of lincRNA, antisense and processed transcript biotypes in case of human, mouse and zebrafish, whereas, for *Drosophila*, total lincRNAs are the representative for lncRNAs and for worm, lincRNAs and ncRNA biotypes together are considered as the lncRNAs.

2.2.3. Data processing

The data processing is a critical phase of the pipeline that involves three important steps - 1) identification and characterization of overlapping features, 2) isolation of sense/antisense (S/AS) pair of coding and lncRNA transcripts 3) mapping of repeat element to the transcripts, while systematically organizing the data into multiple tables for later references. The data processing module of the pipeline requires access to the previously downloaded TranscriptDB and few pre-prepared necessary data tables as input from the data collection module.

2.2.3.1. Identification of overlapping features

The objective of this step is to identify the overlapping transcript pairs which are mapped on opposite stands. For this purpose, the transcripts annotated to plus and minus strands are separated into two groups using the data from TranscriptDB. Subsequently, for these groups, GRanges *(GenomicRanges built upon the IRanges objects with biological semantics corresponding to the transcripts)* data objects are created which mainly contain the genomic coordinate for the transcript start, end position, the strand and corresponding Ensembl transcript ID as identifiers. Taking advantage of the *in-memory* data structure of GRanges objects, the *findOverlap* function from IRanges library is used to collect the overlapping and non-overlapping transcript coordinates. For the overlapping transcripts thus identified, the corresponding Ensembl transcript IDs are extracted and subsequently annotated with other useful information *(Meta-data)* such as Ensembl gene IDs and gene symbols. Additionally, for each overlapping transcript pair, the precise lengths and positions of nucleotides, CDS and exons under overlap are identified and annotated.

More importantly, using the transcript strand information and start, end coordinates, overlapping transcript pairs are characterized based upon the type of overlap they exhibit. Figure 2.2 demonstrates the four possible overlap types between the transcripts on opposite stands - 1) *head-to-head*, where the 5' region of the transcripts on opposite strands are overlapping to each other, 2) *tail-to-tail*, where the 3' region of the transcripts are under overlap 3) *minus inside*, where the transcript on minus strand is completely overlapped by the transcript in the plus strand, lastly 4) *plus inside*, where the transcript on plus strand is in internal overlap with transcript on minus strand.

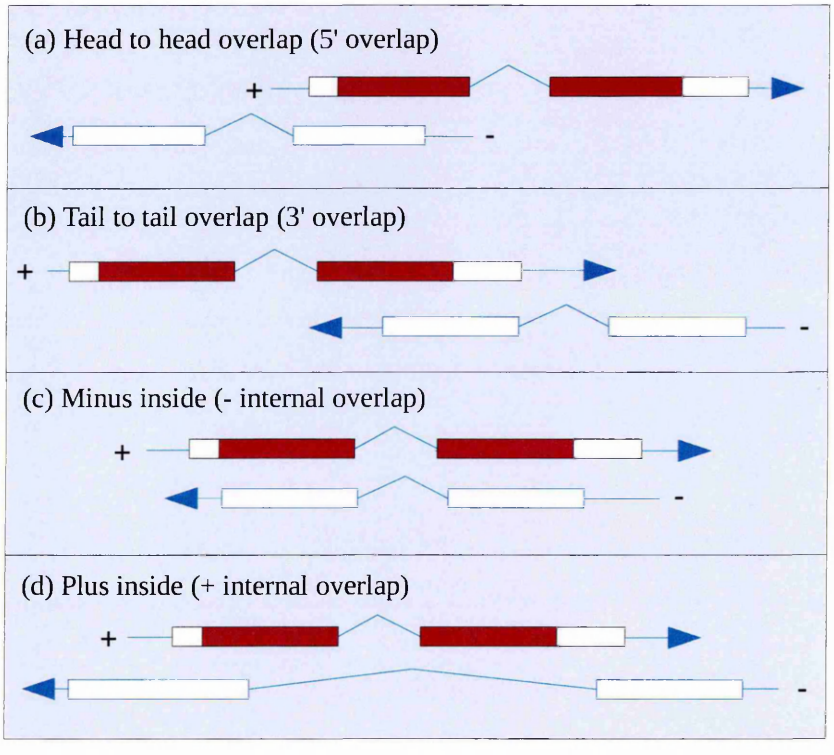


Figure 2.2 | Feature overlap types. The above figure represents all possible overlap types between two transcripts on plus (+) and minus (-) strands where exons are represented by boxes and introns by the connectors between them. The shaded region of the boxes present on (+) strand represents the coding region of mRNA

transcripts while the unshaded regions represents 5' and 3' UTRs. Similarly, the unshaded boxes on (–) strands represents the exons for lncRNA transcripts.

The identification and annotation of transcript overlap type for overlapping transcript pairs is an important step as it enables the isolation of head-to-head overlapping features *(Figure 2.2a)*, that are the center of focus in my study due to their resemblance with the overlap type exhibited by *AS-Uchl1* and *Uchl1* mRNA. These annotations are also be useful for analyzing and comparing the transcript pairs with different types of overlap. The data from this step is systematically organized into data table called *"tx.level" (transcript overlap level information) (table 2.2)* for later reference.

The next step is the identification of the exons participating in the overlap for each transcript pair. This is accomplished similarly with the help of GRanges objects and by using the genomic exon coordinates and exon ranks form TranscriptDB corresponding to the transcripts on plus and minus strand. Followed to this, *findOverlap* function is applied to obtain the overlapping and non-overlapping exons for each transcript pair. Subsequently, the exon overlap lengths are computed and together with other useful information such as the overlap type status form *tx.level* table, are compiled into a new *"ex.level" (exon overlap level information)* data table for later reference *(table 2.2)*.

2.2.3.2 Isolation of S/AS pair of coding and lncRNA transcripts

With the available transcript and exon level overlap information in hand, next step is the isolation of S/AS transcript pairs where, one of the transcript is an mRNA and the other a lncRNA. This is accomplished by annotating the previously downloaded transcript and gene

biotype information to the overlapping transcript pairs in *ex.level* table using Ensembl transcript and gene IDs as identifiers. Another crucial step at this stage is to identify if the translation initiation sites (TIS/ATG) for the overlapping coding transcripts lie within the overlap region. This is identified with the help of pre-computed exon overlap length, CDS start, end coordinates and the transcript strand information. Annotation of ATG overlap status to the transcript pairs is an important step, as it is used for the investigation of the modular nature of ASlncRNAs in later steps. The isolated S/AS pair of coding and lncRNA transcripts along with the useful information such as ATG overlap status are stored separately in *"sas.pc_nc"* (S/AS pair of protein-coding and lncRNA transcripts) *(table 2.2)* data table for later reference.

2.2.3.3. Mapping of repeat elements

For mapping repeat elements to the transcript exons, GRanges objects are created for repeats, using previously downloaded RepeatMasker data and for exons, using the exon information from the TranscriptDB. With the available genomic coordinates for repeat elements and exons in GRanges data structure, once again *findOverlap* function is implied for the identification of the repeat elements overlapping to exons for each transcript. Along with the repeat mappings, precise number of exonic nucleotides covered by each repeat element and their orientation with respect to the transcripts are systematically compiled into the *"repeat.level" (repeat overlap level information) (table 2.2)* table which serve as crucial information for later analysis.

Finally, the previously generated *sas.pc_nc* table is merged together with *repeat.level* table to generate a comprehensive resource for S/AS pair of coding and lncRNA transcripts in the

trancriptome, which contains all the necessary information such as overlap type, ATG overlap, repeat content and specific orientation in ASlncRNAs, length of repeat overlap etc. This information is separately organized into a table addressed as *"sas.pc_nc.repbase" (S/AS transcript pairs with repeat annotations for AS counterpart) (table 2.2).*

Table name	Data information
transcriptome_ statistics	general statistics of the transcriptome data (eg; No. of transcripts, No. of genes belonging to different biotypes)
ge_tx_info_df	gene names, gene symbols, transcript ids, transcript start, end coordinates, strand, biotypes, cds length, peptide ids
tx_ex_info_df	transcript ids, exon ids, exon ranks, exon start, end coordinates, strand, cds start, end coordinates
go.anno.db	gene ids, go annotations (cellular component, molecular function, biological process)
go.dualloc.anno.db	gene ids, dual location annotations considering nucleus, cytoplasm and mitochondrion
homolog.genes	homolog genes considering all species under analysis
tx.level	overlapping transcripts present in opposite strand, overlap type annotations *(eg; head-to-head, tail-to-tail etc)*
ex.level	overlapping exons for the overlapping transcripts, TIS for coding transcripts, cds start, end coordinates, overlapping exon rank, TIS overlap status
repeat.level	transcript ids, annotation of overlapping repeat class, family and elements, exon rank contain repeats, repeat orientations with respect to transcripts
sas.pc_nc	S/AS pair of coding and ASlncRNA transcripts, TIS overlap status, overlap length, exons in overlap, overlap type annotations
sas.pc_nc.repbase	S/AS pair with repeat overlap annotations for ASlncRNA transcripts, repeat overlap length, repeat overlap type annotation, repeat orientation, transcript overlap type annotation, TIS overlap status
noas	list of transcripts that do not overlap with other transcripts, transcript, exon ids
transcript categories for TE coverage analysis	transcripts annotated with repeats, repeat overlap length
gene catagories for functinal enrichment analysis	genes ids classified based up on overlap type annotations, TIS overlap, SINE orientations in ASlncRNAs
complete.pc.norm	gene ids, transcript ids corresponding to the representative transcript isoform containing longest coding region (cds length), peptide ids
norm_pep.seq	protein sequences corresponding to the representative longest coding transcript for each coding gene

Table 2.2 | Data tables generated by pipeline. Table contains the names of the main tables generated by the pipeline (left) and the information present in each of them *(right)*

2.2.4. Data analysis and representation

2.2.4.1. Determination of repeat content

In order to infer repeat content of coding and non-coding genes in the transcriptome, the data analysis module of the pipeline makes use of the previously generated *repeat.level* table and computes the percentage of total protein-coding and lncRNA genes *(Table 2.1)* that contain specific repeat class within the exons of their transcripts. For this, a minimum overlap of at least 10 nucleotide between the exon and the repeat element is taken into consideration with reference to a previously published study by *Kapusta et al., 2013*.

2.2.4.2. Classification and nomenclature of the transcripts.

In order to study the contribution of TEs to the exonic sequence of ASlncRNAs in contrast to rest of the transcripts in the transcriptome, there is a need of systematic categorization of lncRNA considering their head-to-head overlap against the genomic span (i.e. exon) of a protein-coding locus on the opposite strand. Such a categorization could yield multiple transcript classes, each with specific "characteristics", thus demanding a formal nomenclature *(Table 2.3)*. Here, with "characteristics", I meant to address the "biotype" of the transcripts belonging to each gene. It is important to bear in mind that a single coding gene could have multiple transcript isoforms with different biotypes (coding/non-coding) and each of them could overlap to another transcript on the opposite strand. When a non-coding transcript isoform corresponding to a coding gene, overlap to a coding isoform of a coding gene in opposite strand, they could together form a S/AS transcript pair similar to that of *Uchl1* mRNA and *AS-Uchl1 lncRNA*. Therefore, such non-coding transcripts could not be ignored just because they are the isoforms of a protein-coding gene, but rather are needed to be considered as ASlncRNA yet categorized into a separate class, so that they could be

distinguished from ASlncRNA transcript isoform a non-coding gene. It is important to classify the transcripts in this manner, because this would give us an opportunity to perform several comparative analysis to determine the contribution of TEs to their sequences and to understand how different or similar they could be from each other in terms of their sequence evolution. Before we proceed any further, it is very important to get familiarized with the nomenclature that is used throughout the thesis to address different transcript categories analyzed in the study. The names for each class is tagged with a short string which contains the information regarding the biotypes for the overlapping sense and antisense transcripts (isoforms) and genes. The assigned names for each transcript category could be difficult to understand at a first read-through, but it is also a compact and intuitive nomenclature that explains the overlap type exhibited by each transcript belonging to a gene. It is worth while to read carefully the description for each category name mentioned in *table 2.3*, because this would help in easier interpretation several comparative analysis discussed further in this thesis.

No.	Categories	Description
1	NCall	All long non-coding transcripts
2	NCnc-PCpc	Non-coding transcripts antisense to coding transcripts
3	NCpc-PCpc	Non-coding isoforms of coding genes antisense to coding transcripts
4	PCpc-PCpc	Coding isoforms of coding genes antisense to coding transcripts
5	NCnoas	Non-coding transcripts with no antisense evidence
6	PCall	All coding transcripts
7	PCpc-NCnc	Coding transcripts antisense to non-coding transcript
8	PCpc-NCpc	Coding transcripts antisense to non-coding isoform of coding genes
9	PCnoas	Coding transcripts with no antisense evidence
10	NCpcnoas	Non-coding isoforms of coding genes with no antisense evidence
11	PCpcnoas	Coding isoforms of coding genes with no antisense evidence

Table 2.3 | Transcript categories. The above table represents the schema for indicating different transcript categories that are automatically generated by the TEs coverage enrichment module of the pipeline. Here, the *NCall* and *PCall* categories represent total non-coding and coding transcripts respectively. Here, the nomenclature for the categories are designed considering different types of transcript isoforms corresponding to a single gene. The category names containing the "-" symbol represents the transcripts that take part in sense/antisense overlap, where, the portion of the name before "-" specify the antisense transcript and the portion after, represents the transcripts in sense orientation. Additionally, each sub-portion of the category names have two uppercase and two lowercase letters which indicates specific biotypes for transcript and its relative gene respectively. This distinction is important because many coding genes also transcribe non-coding isoforms. For example, NCpc-PCpc indicates a S/AS pair in which the antisense is a

non-coding isoform of a coding gene and the sense is the coding isoform of a coding gene. In the same way NCnc-PCpc indicates a S/AS pair in which the antisense is a non-coding isoform of a non-coding gene and the sense is the coding isoform of a coding gene. The measures of TEs coverage percentage are always referred to the antisense transcript, i.e. on the transcript that appear before the "-" in the notation.In addition to the indicated classes in the thesis I have also used the additional 2 super-categories: ASlncRNAs made by the union of *NCnc-PCpc* and *NCpc-PCpc* and noASlncRNAs made by the union of *NCnoas* and *Ncpcnoas.*

2.2.4.3. Computation of TEs coverage enrichment across different transcript categories

For each of the transcript categories described in *table 2.3,* the percentage of sequence covered by TEs is computed considering the total number of nucleotides contributed by each specific TE class. The computed percentages are subsequently represented as simple comparative stacked bar charts for better interpretation. However, to analyze if the coverage of specific TEs are significantly higher among ASlncRNAs *(represented by NCnc-PCpc and NCpc-PCpc transcript categories)* in contrast to noASlncRNAs *(represented by NCnoas and NCpcnoas transcripts)* there is a need of statistical comparative analysis, considering both these transcript groups. For this, the TEs coverage enrichment module of the pipeline performs a randomization analysis, wherein 1000 random samples of noASlncRNAs are generated with the sample size *n,* same as the total number of ASlncRNA transcripts. The main motive behind generating the random samples was to compare the mean of the percentage of sequence covered by each TEs in the random samples, with that of the actual percentage of sequence covered by TEs in ASlncRNA transcripts. For this comparison, I decided to use the Z-test which is an appropriate test because, the sample size for the samples under comparison are

large in this analysis. The Z test is a statistical procedure used to test an alternative hypothesis against a null hypothesis. In this case, the **null hypothesis** is - *the percentage of sequence covered by TEs among ASlncRNA is not different from that of the mean TEs coverage in random samples of noASlncRNAs,* whereas, the **alternative hypothesis** is - *the percentage of sequence covered by TEs in ASlncRNAs is either significantly higher or lower than the mean TEs coverage observed in random samples of noASlncRNAs*. Here, the Z-test builds upon the Z-score which is a measure of how many standard deviations below or above, the TEs coverage in ASlncRNAs is, from that of the mean TEs coverage in noASlncRNA random samples.

The formula for calculating Z-score is

$$z = (X - \mu) / \sigma$$

where, z is the Z-score, X is the TEs coverage observed among ASlncRNAs, μ is the mean of TEs coverage in 1000 samples of noASlncRNAs and finally σ is the observed standard deviation in TEs coverage for the population of noASlncRNA random samples. The obtained Z-score could then be placed in the normal distribution to determine whether or not to reject the null hypothesis. However, the probability of falsely rejecting a null hypothesis is determined by p-values which could be calculated considering a two sided test *(in this case),* either manually by reading a Z-score table (also referred as standard normal table), or automatically using the *pnorm()* function in R ("pnorm" stands for "probability normal distribution").

The *pnorm()* function in R calculates the cumulative distribution function (cdf) i.e,

$$F(x) = P(X <= x)$$

Where, X is normal and x is the test statistic. The above expression stands good for a one-sided test or more specifically, the lower-tailed test, where the distribution function X is evaluated at x, to calculate the probability that X will take a value less than or equal to x. Hence, here the p-value would be equal to ***cdf(x)***. Similarly, If the probability of X being more than or equal to x were to be tested *(an upper-tailed test)*, then the p-value would be equal to ***P(X >= x) or 1 – cdf(x)***. Since in my analysis, I am interested to determine if the TEs coverage in ASlncRNAs is not equal to the mean of TEs coverage in 1000 samples of noASlncRNA *(a two tailed test)*, the formula for calculating p-value can be established as

$$p\text{-}value = P(X >= x) + P(X <= x)$$

Which is simply equals to, ***2 * P(X >= |x|)*** or can also be written as, ***2 * (1-cdf(|x|))***. In R this can be easily calculated using *pnorm()* as, ***p-value = 2 * pnorm(-abs(z))***, where z is the Z-score.

Usually, a very high or a very low Z scores (fitting in the tails of the normal distribution) are associated with very small p-values, indicating a low chance of falsely rejecting the null hypothesis. This here would mean that the percentage of specific TEs coverage among ASlncRNAs is significantly higher or lower in comparison to their coverage in noASlncRNA sequences. In general, the p-value threshold for significance is set to 0.05, this means that there is a 5% of chance that the result of the test is a false positive (Dorey, 2010). In other words, although based on the TEs coverage enrichment analysis, the TEs coverage in

ASlncRNA is identified as significantly different from the TEs coverage in noASlncRNAs, in reality there is no such difference. If the 5% of chance of the result being false positive is acceptable for one single test, then for example, 1000 such tests could result into the discovery of 50 false positive results, just by chance. This is known as the multiple comparisons or **multiple testing problem** in statistics.

To overcome such problem of multiple testing, I have implemented the FDR (False discovery rate) method (Benjamini & Hochberg, 1995) which assigns an adjusted p-value keeping into account the total number of statistical tests performed in a single analysis with the effect of reducing in practice the p-value threshold from 5% to a more reasonable value. The FDR method is designed to control the number of false discoveries by controlling the rate of type I errors, i.e., accepting a false hypothesis as correct (Benjamini & Yekutieli, 2001). Bonferroni correction is another similar method to check on the multiple testing problem, however it is considered as too conservative in terms of reducing the number of true discoveries, while reducing the number of false positives (Glickman, Rao, & Schultz, 2014).

2.2.4.3.1. Multiple testing correction with FDR method

Benjamini and Hochberg defined the FDR as follows

FDR = expected proportion of erroneous rejections among all rejections

Based on the notations used in *table 2.4* this can be shown as,

$$\mathrm{FDR} = \frac{V}{V+S} = \frac{V}{R}$$

Here, FDR *(also denoted as q or q* value by Benjamini and Hochberg),* is an unobserved value random variable, as we do not know *V* and *S*. Hence FDR can also be defined as the expectation of *(V/R)* like this

$$\mathrm{FDR} = E\left(\frac{V}{R}\right)$$

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Table 2.4 | Number of errors committed when testing *m* null hypotheses. The above table summarizes the notations used by Benjamini and Hochberg. Here ***m*** represents the hypotheses that are assumed to be known in advance; $\boldsymbol{m_0}$ are total true hypotheses; ***R*** is an observable random variable (Total no. of all rejected null hypothesis) ; *U, V, S* and *T* are the unobservable (unknown) random variables, where ***U*** is the no. of true null that are not rejected, ***V*** is the no. of true null that are rejected, ***T*** is the no. of false null that are not rejected and ***S*** is the no. of false null that are rejected. (The above table is taken from *Benjamini & Hochberg, 1995*).

The method of Benjamini & Hochberg can be described using the following steps

- Lets use q to denote FDR that is considered tolerable. It is typically set to .05 to ensure that the chances of falsely rejecting a true null hypothesis is fairly small
- Let $p_1, p_2, p_3, p_4, \ldots\ldots, p_m$ be the p-values of m tests performed, that are ordered from smallest to highest
- Let $k = 1, 2, 3, 4, \ldots\ldots, m$ be the indices of the ordered p-values
- Calculate a threshold value for each p-value using, $(k * q)/ m$
- Compare each p-value $p_{(k)}$ against its corresponding threshold value $(k*q)/m$ using the following expression

$$\hat{k} = \max\left\{k: p_{(k)} \leq \frac{k \cdot q}{m}\right\}, \quad k = 1, 2, \ldots, m$$

- Finally, if $\hat{k} \geq 1$ then reject the hypotheses that correspond to $p_1, p_2, \ldots, p_k$ and fail to reject the hypotheses that correspond to the rest of the p-values (Benjamini & Hochberg, 1995; Jack Weiss, 2003).

The FDR method based adjusted p-values can be extracted using the expression used to compare each p-values against the threshold value i.e,

$$p_{(k)} \leq \frac{k \cdot q}{m}$$

Here, q is the FDR, hence solving for q in the above expression gives, $q = m * p_{(k)} / k$. Using this, the adjusted p-values can be determined with following expression

$$\tilde{p}_{(i)} = \min_{k \in \{i,\ldots,m\}} \left\{ \min\left(\frac{m}{k} p_{(k)}, 1 \right) \right\}$$

Here, $\tilde{p}_{(i)}$ is the adjusted p-value. When the adjusted p-value is less than q, then the null hypothesis can be rejected (Benjamini & Hochberg, 1995; Jack Weiss, 2003).

In my analysis, I have implemented *p.adjust()* function in R for adjusting the p-values using the Benjamini & Hochberg's FDR method *(here and elsewhere in this thesis)*. The function offers several other methods for the p-adjustment, wherein I have selected FDR using the "method" parameter of the function. (Here, is the link to read more about the *p.adjust()* function in R https://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html)

2.3. Results and discussions

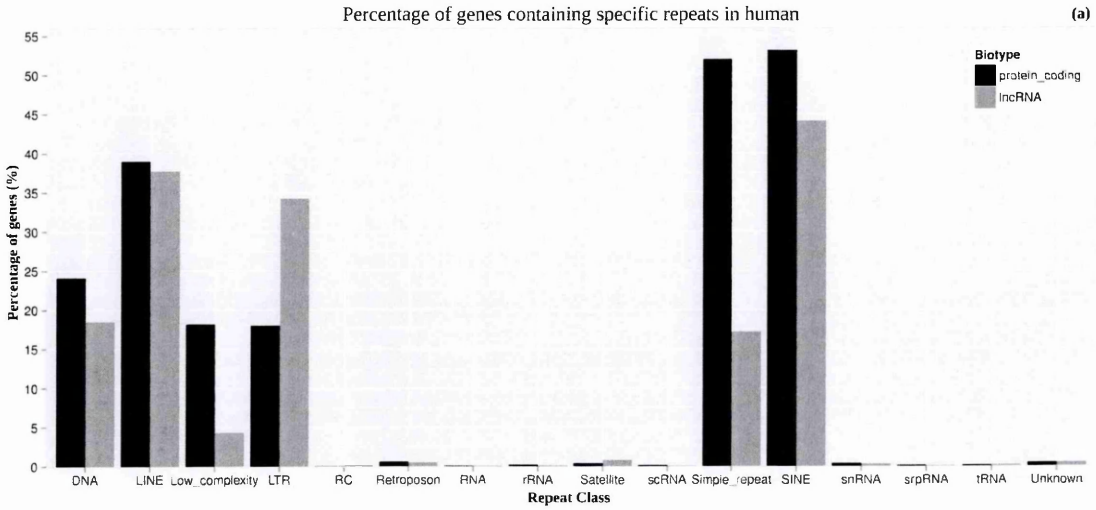
2.3.1. Percentage of protein-coding and lncRNA genes containing specific repeats

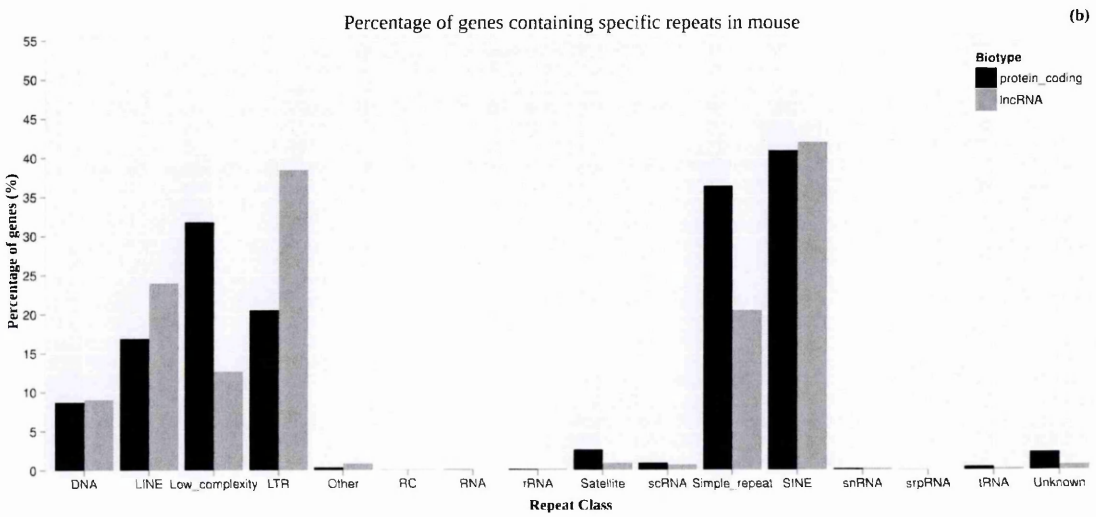
The percentage of coding and lncRNA genes containing different repeat elements *(Figure 2.3)* reflect several well known facts about the genomes of all five species analyzed in the study. For example, human and mouse genomes contain a higher density of TEs, where retrotransposons account for a major fraction of all interspersed repeats. In contrast to human and mouse, zebrafish genome is dominated by DNA transposons, whereas other TEs such as LINEs are known to have experienced continual turnover among teleost fish genomes (Duvernell, Pryor, & Adams, 2004; Kapusta et al., 2013; Krasnov et al., 2005). Finally, fruit-fly genome is predominantly covered by simple repeats such as CAG/CTG trinucleotides, that are less frequent among humans and *C. elegans* genome sequences (Katti, Ranjekar, & Gupta, 2001).

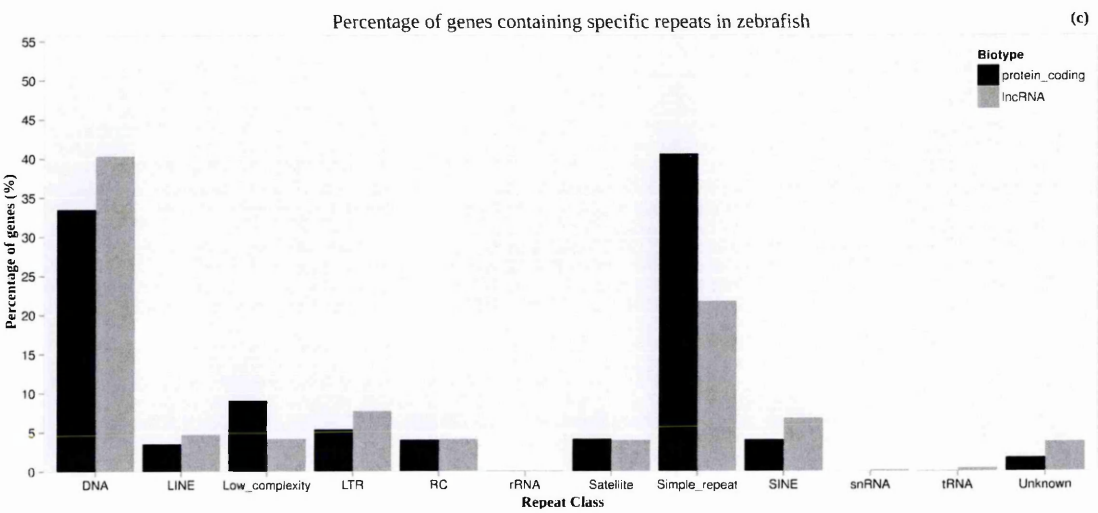
Other interesting observations that could be drawn from figure 2.3 are as follows -1) Higher fraction of genes contain TEs in case of human and mouse (mammals), whereas zebrafish, a non-mammalian vertebrate contain lower fraction of genes with TEs in contrast to mammals, but higher than invertebrates such as fruit-fly and worm *(Figure 2.3 c-e).* This differences in the percentage of genes containing TEs among different species could be explained by the phenomenon of TE exonization. In a comparative genomic study *Sela et al.*, identified that the rate of TEs exonization widely vary among different species, where mammals show a higher magnitude of TE exonization followed by other vertebrates in comparison to invertebrates. They also identified the abundance of TEs in intronic sequences are much higher among vertebrates than in invertebrates, suggesting that TEs located within long introns provide a possibility for testing new exons through the process of exonization in vertebrates (Sela, Kim,

& Ast, 2010). 2) The percentage of genes containing TEs such as LINEs and SINEs are slightly lower than what is observed in human *(Figure 2.3 a, b).* Such a difference could be explained by the fact that the rodent lineages have undergone greater rate of molecular evolution and sequence substitutions (W.-H. Li, 1997), which makes the recognition of ancestral TEs such LINEs *(L2)* and SINEs *(MIR)* more difficult in rodents. The transposition activity of the ancestral L2 and MIR *(Mammalian-wide interspersed repeat)* are known to be ceased before the split of primate-rodent lineages and have undergone residual movements since then, suggesting for a selective constraint on mammalian TEs since at least the divergence of humans and mice (Silva et al., 2003; Mouse Genome Sequencing Consortium, 2002).

In summary, this exploratory analysis reveals the percentage of coding and lncRNA genes containing specific repeats among all analyzed species. The difference between the percentage of coding and lncRNA genes containing specific TEs is comparable, if not significantly different among all species. A specific trend of TEs distribution could be observed between the lncRNA and coding genes, for example, non-LTR TEs are present in slightly higher percentage of lncRNA genes in mouse while this is exactly opposite in human. In case zebrafish, DNA tranposons are present in higher fraction of lncRNAs than coding genes while simple repeats are present in higher fractio of coding genes. Similarly, a higher fraction of lncRNAs contain low-complexity repeats in fly and worm etc. However, to gain a better insight of the contributions of TEs to the exonized region of the transcriptome it is important to examine the TEs coverage (Percentage of nucleotide covered by TEs) instead of TEs content (Percentage of genes containing TEs). As a consequence I next analyzed the TEs coverage among the transcript categories explained in *table 2.3.*


Percentage of genes containing specific repeats in human
(a)
Biotype
protein_coding
lncRNA
Percentage of genes (%)
0
5
10
15
20
25
30
35
40
45
50
55
DNA
LINE
Low_complexity
LTR
RC
Retroposon
RNA
rRNA
Satellite
scRNA
Simple_repeat
SINE
snRNA
srpRNA
tRNA
Unknown
Repeat Class


Percentage of genes containing specific repeats in mouse
(b)
Biotype
protein_coding
lncRNA
Percentage of genes (%)
0
5
10
15
20
25
30
35
40
45
50
55
DNA
LINE
Low_complexity
LTR
Other
RC
RNA
rRNA
Satellite
scRNA
Simple_repeat
SINE
snRNA
srpRNA
tRNA
Unknown
Repeat Class


Percentage of genes containing specific repeats in zebrafish
(c)
Biotype
protein_coding
lncRNA
Percentage of genes (%)
0
5
10
15
20
25
30
35
40
45
50
55
DNA
LINE
Low_complexity
LTR
RC
rRNA
Satellite
Simple_repeat
SINE
snRNA
tRNA
Unknown
Repeat Class

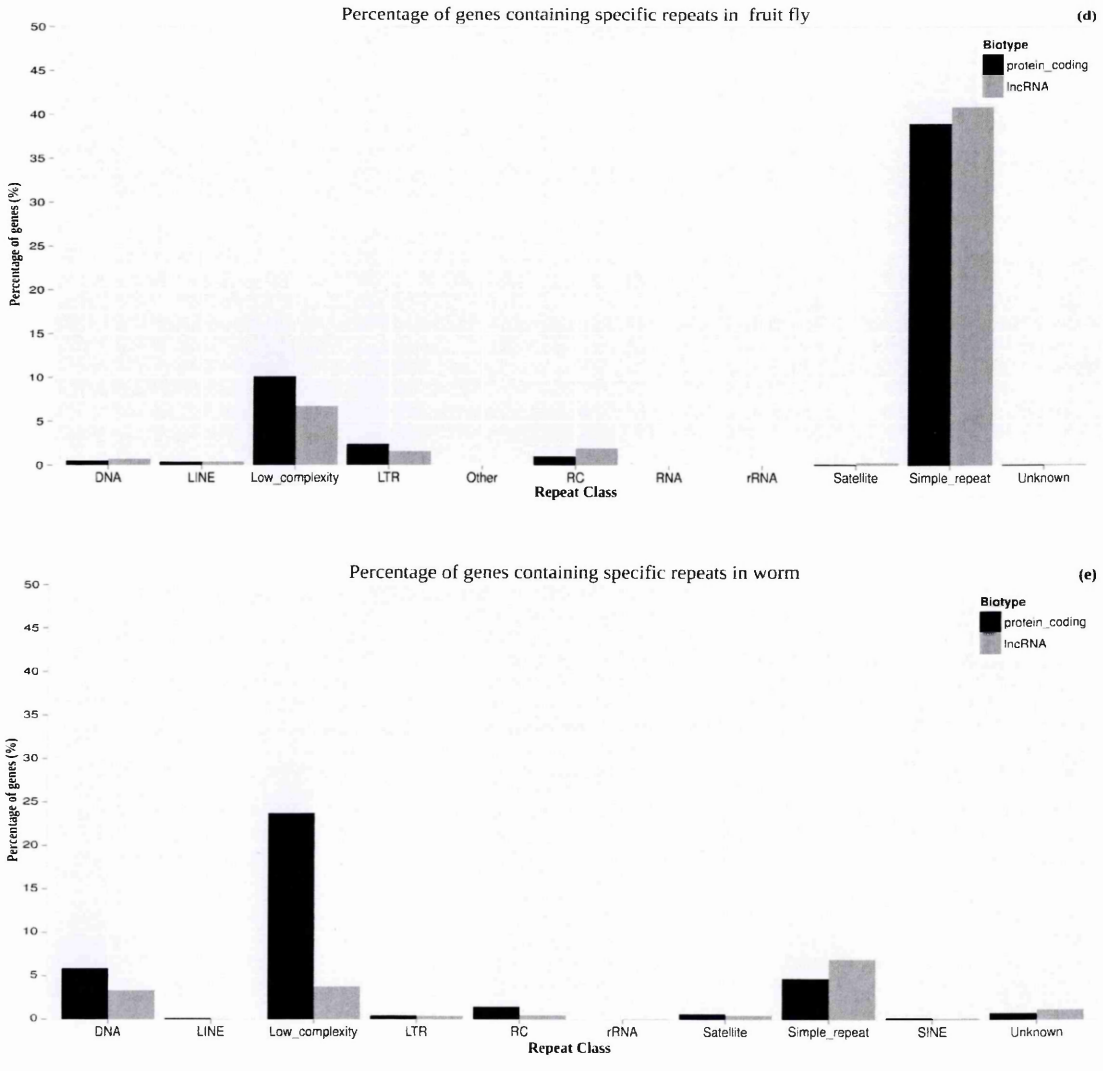


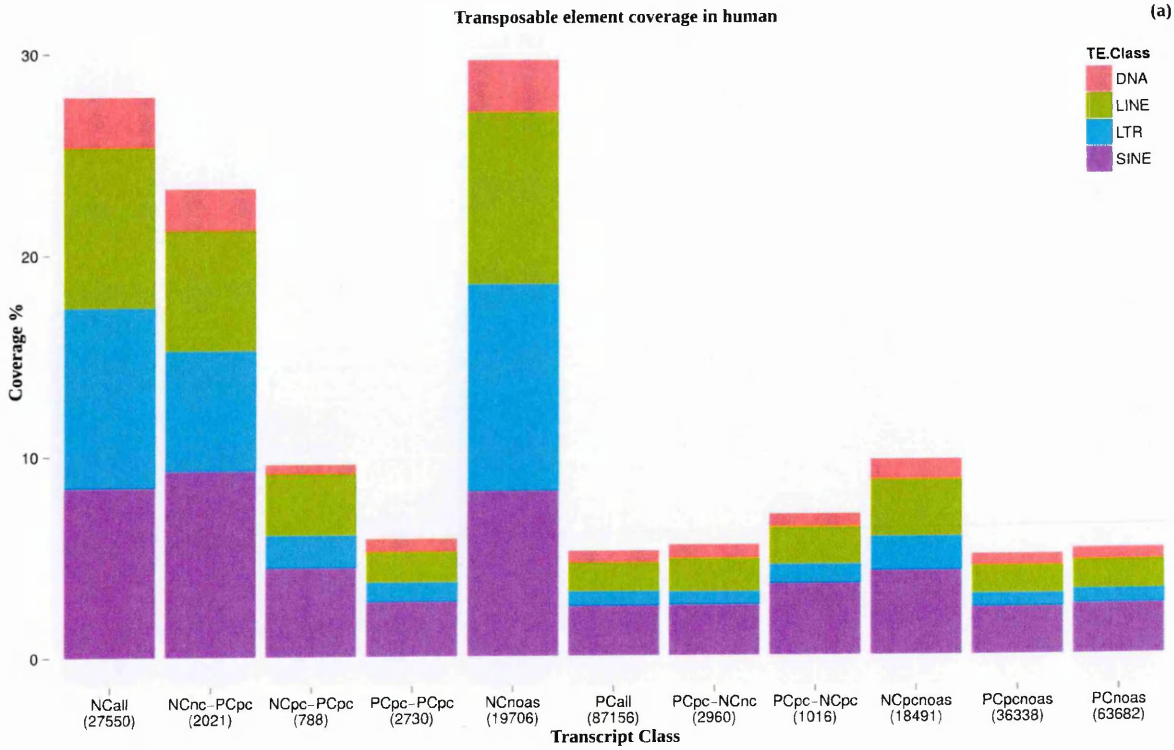
Figure 2.3 | Percentage of genes containing repeat elements. The above charts represents the percentage of total coding and lncRNA genes *(y-axis)* containing specific repeat elements annotated by RepeatMasker *(x-axis)* for (a) human, (b) mouse, (c) zebrafish, (d) fruit-fly and (e) worm respectively.

2.3.2. TEs coverage analysis

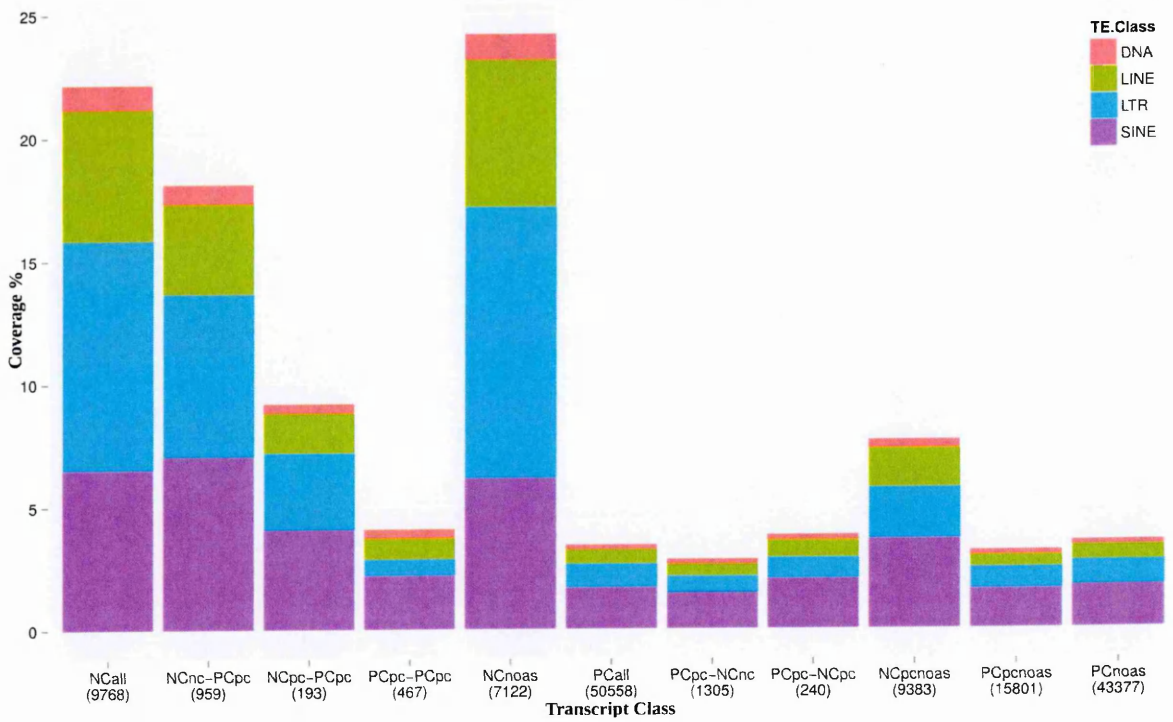
The TEs coverage analysis is performed across different groups of transcripts *(Table 2.3)* as described in section ***2.2.4.3***. The computed percentage of TEs are then represented as stacked bar chart for all the analyzed species as shown in *figure 2.4*. One of the easily accountable observations from the charts for all species is that the transcript categories containing lncRNAs show a higher fraction of total TEs coverage with respect to coding transcript categories. This implies that the majority of lncRNAs are TEs associated in all species.

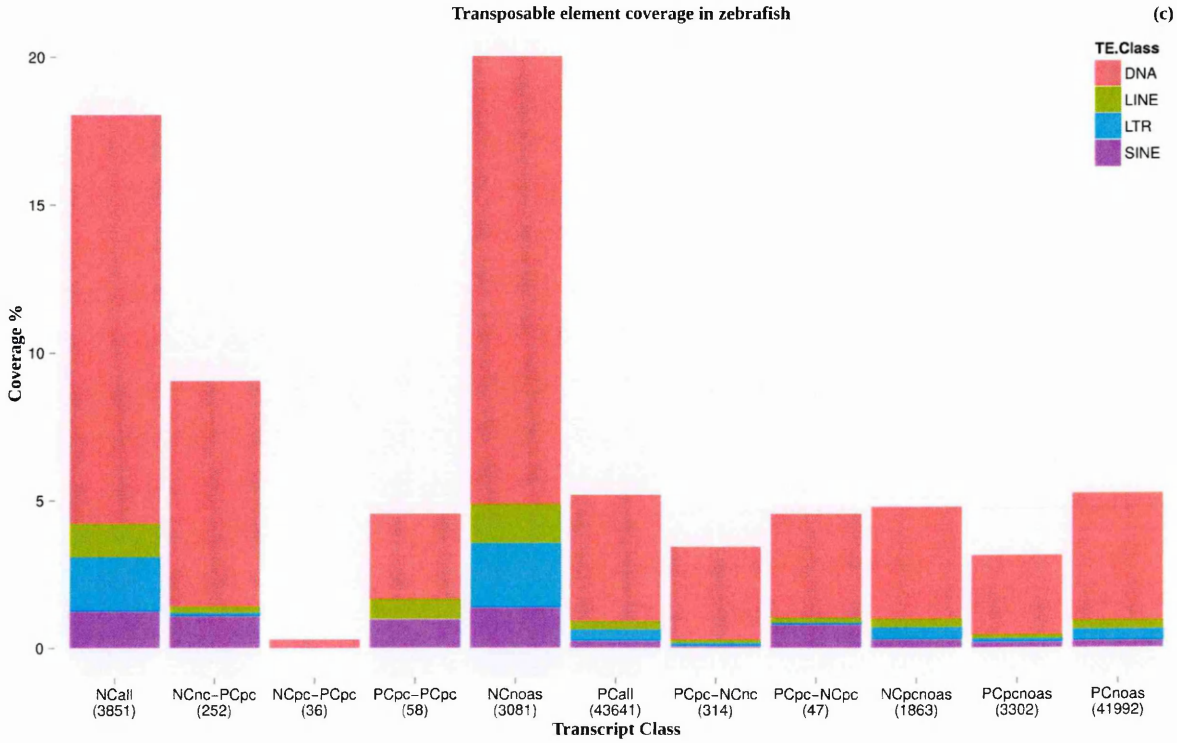
Further, in order to analyze if the observed differences in the TEs coverage between ASlncRNA and noASlncRNA sequences are statistically significant, I performed an enrichment analysis by generating 1000 random samples of noASlncRNAs followed by determining the TEs coverage in each of them. The main motive of doing this was to compare the TEs coverage observed in ASlncRNA sequence against the mean TEs coverage observed in the 1000 eandom samples of noASlncRNA. For this comparison, I performed a Z-test (discussed in detail in the methods section ***2.2.4.3).*** The result of the enrichment test is shown in *Table 2.5*.

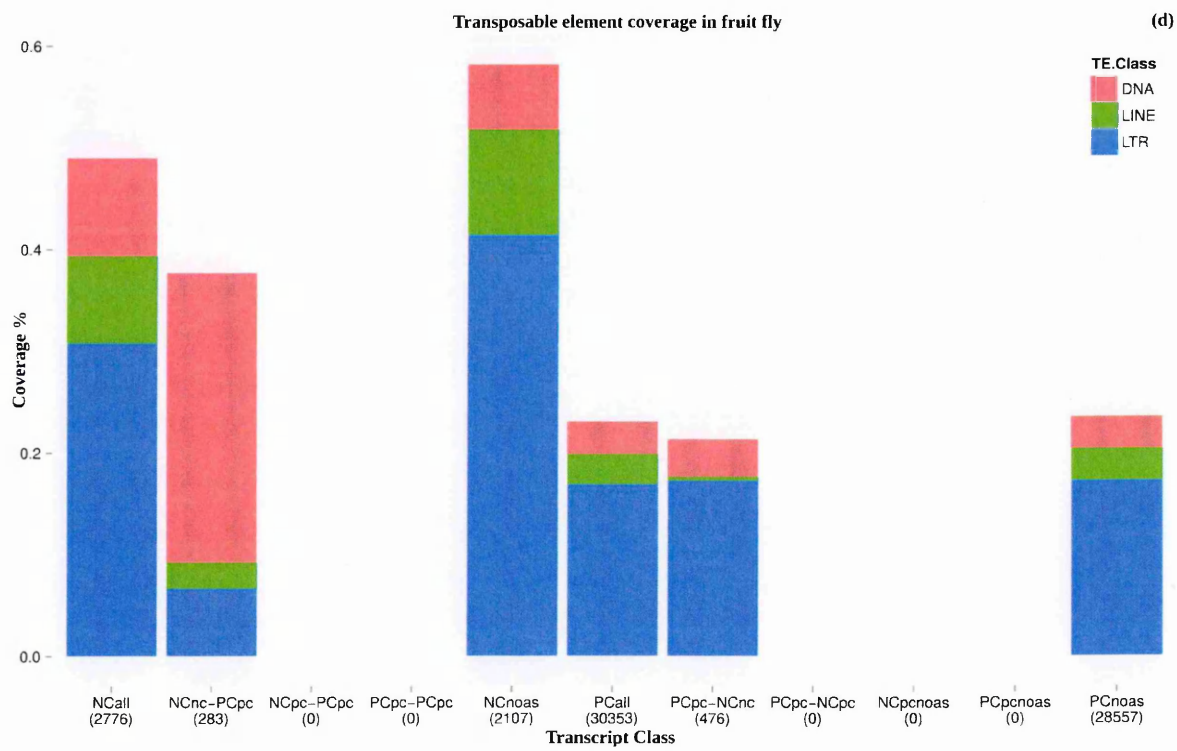
(a)


Transposable element coverage in human
TE.Class
DNA
LINE
LTR
SINE
Coverage %
30
20
10
0
NCall (27550)
NCnc-PCpc (2021)
NCpc-PCpc (788)
PCpc-PCpc (2730)
NCnoas (19706)
PCall (87156)
PCpc-NCnc (2960)
PCpc-NCpc (1016)
NCpcnoas (18491)
PCpcnoas (36338)
PCnoas (63682)
Transcript Class

(b)


Transposable element coverage in mouse
TE.Class
DNA
LINE
LTR
SINE
Coverage %
25
20
15
10
5
0
NCall (9768)
NCnc-PCpc (959)
NCpc-PCpc (193)
PCpc-PCpc (467)
NCnoas (7122)
PCall (50558)
PCpc-NCnc (1305)
PCpc-NCpc (240)
NCpcnoas (9383)
PCpcnoas (15801)
PCnoas (43377)
Transcript Class


Transposable element coverage in zebrafish
(c)
TE.Class
DNA
LINE
LTR
SINE
Coverage %
20
15
10
5
0
NCall (3851)
NCnc-PCpc (252)
NCpc-PCpc (36)
PCpc-PCpc (58)
NCnoas (3081)
PCall (43641)
PCpc-NCnc (314)
PCpc-NCpc (47)
NCpcnoas (1863)
PCpcnoas (3302)
PCnoas (41992)
Transcript Class


Transposable element coverage in fruit fly
(d)
TE.Class
DNA
LINE
LTR
Coverage %
0.6
0.4
0.2
0.0
NCall (2776)
NCnc-PCpc (283)
NCpc-PCpc (0)
PCpc-PCpc (0)
NCnoas (2107)
PCall (30353)
PCpc-NCnc (476)
PCpc-NCpc (0)
NCpcnoas (0)
PCpcnoas (0)
PCnoas (28557)
Transcript Class

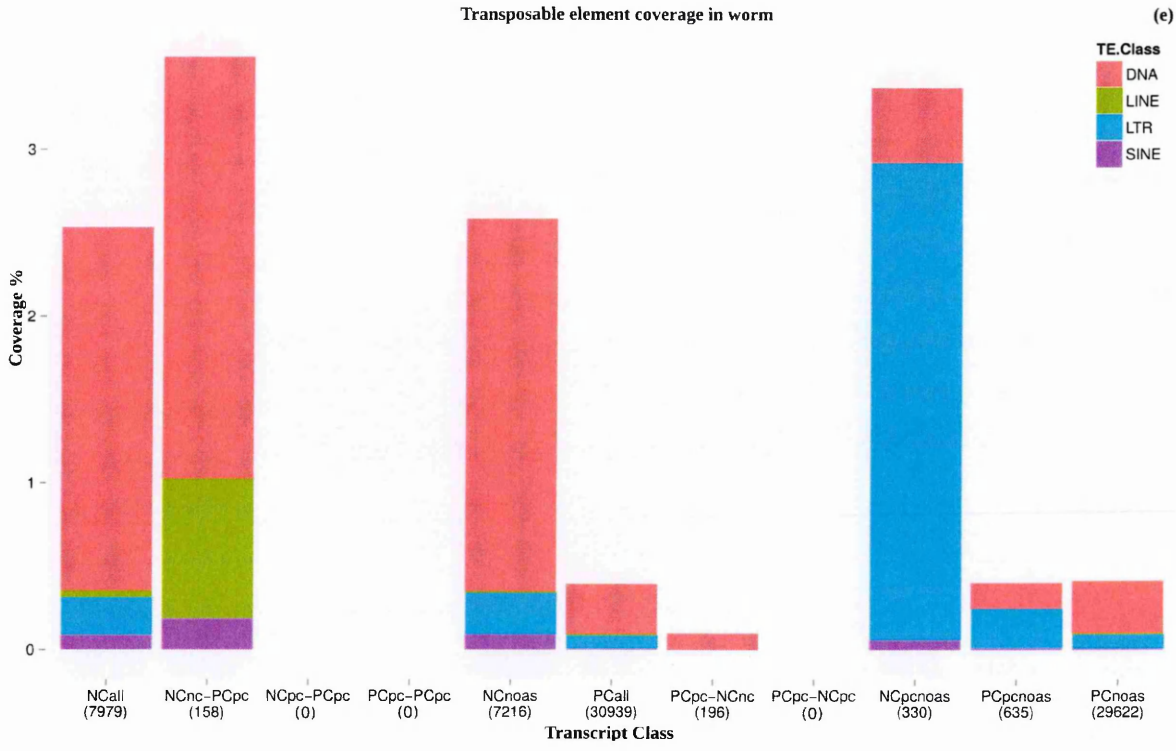


Figure 2.4 | TEs coverage. The above charts show the percentage of total transcript nucleotides represented by the different TE classes *(y-axis)* among the different transcript categories. The total number of transcripts in each category is indicated below the category name within brackets *(x-axis)*. Charts are for (a) human, (b) mouse, (c) zebrafish, (d) fruit-fly and (e) worm respectively.

Based on the TEs coverage enrichment analysis, one the most interesting observation that can be noted is that, in case of both human and mouse the ASlncRNA sequences *(represented by NCnc-PCpc, NCpc-PCpc transcript categories)* are significantly enriched for SINE elements, where SINEs account for 1.3 and 1.4 fold higher fraction of sequence coverage in contrast to noASlncRNAs *(represented by NCnoas, NCpcnoas transcript categories)* in human and mouse respectively *(Figure 2.4 a,b; Table 2.5).* At the same time, the coverage of other TEs are marginally lower or equal in ASlncRNAs in comparison to noASlncRNAs. This implies that the ASlncRNAs in human and mouse tend to retain SINE elements but not other class of TEs, which therefore might have associated functional implications.

In case of the remaining species, SINE repeats are the subordinate TEs that covers only a small fraction of sequences in different transcript categories. In zebrafish, the DNA transposons are the dominant class of TEs. However, when the ASlncRNAs are compared against noASlncRNAs, DNA transposons show a 2 fold lower coverage in ASlncRNAs. Apart from DNA transposons, LTRs are also significantly depleted in ASlncRNA sequences in contrast to noASlncRNAs *(Figure 2.4 c; Tabl2 2.5).* In case of fruit-fly, there are no RepeatMasker annotated SINE TEs and the total number of ASlncRNAs are relatively fewer in comparison to the other analyzed species. The ASlncRNAs in fruit-fly are only represented by *NCnc-PCpc* group of transcripts *(Figure 2.4 d) (abbreviations described in Table 2.3)* that are significantly enriched for DNA transposons with a 4 fold higher fraction of ASlncRNA sequence coverage in respect to noASlncRNAs *(Table 2.5).* Lastly, in case of worm, the ASlncRNA sequences are significantly enriched for LINE elements that shows a 2 fold higher fraction of ASlncRNA sequence coverage in contrast to noASlncRNAs *(Figure 2.4 e; Table 2.5).*

Species	TE Class	ASlncRNA (Coverage %)	noASlncRNA (Coverage %)	Z-score	p-value	p-value (Adjusted)
Human	DNA	1.7392813588	1.8648040761	-0.8360114851	4.03e-001	4.03e-001
	LINE	5.3562989215	5.9770507143	-1.7328841122	8.31e-002	1.11e-001
	LTR	5.0514940707	6.4216281385	-3.6069075857	3.10e-004	6.20e-004
	SINE	8.2493596793	6.4104756814	6.5940125792	4.28e-011	1.71e-010
Mouse	DNA	0.7462595442	0.7049867776	0.3399932393	7.34e-001	7.34e-001
	LINE	3.4447726421	3.7614834363	-0.5705972397	5.68e-001	7.34e-001
	LTR	6.2194197211	6.5464771888	-0.5620952894	5.74e-001	7.34e-001
	SINE	6.7398415778	4.8941604601	5.8834577157	4.02e-009	1.61e-008
Zebrafish	**DNA**	7.6194196288	15.172960399	-3.8197989481	1.34e-004	5.34e-004
	LINE	0.2142056122	1.3294879188	-2.3625774885	1.81e-002	2.42e-002
	LTR	0.1255832903	2.1495509477	-3.0827803856	2.05e-003	4.10e-003
	SINE	1.080688314	1.3807633528	-0.7814564329	4.35e-001	4.35e-001
Fruit-fly	**DNA**	0.2847171198	0.0665105781	2.8564648975	4.28e-003	1.29e-002
	LINE	0.0259131941	0.1069165121	-0.9591275381	3.37e-001	3.37e-001
	LTR	0.0662590532	0.4158329369	-1.5138722953	1.30e-001	1.95e-001
	SINE	NA	NA	NA	NA	NA
Worm	DNA	2.5283986808	1.7603969136	0.741654808	4.58e-001	7.25e-001
	LINE	0.8366923171	0.0053877181	20.7458080837	1.34e-095	5.35e-095
	LTR	0	0.8690421032	-0.2616876683	7.94e-001	7.94e-001
	SINE	0.1872887912	0.0794166563	0.6071084676	5.44e-001	7.25e-001

Table 2.5 | TEs class coverage enrichment. Above table contains the Z-scores and p-values generated from the enrichment analysis for TEs class coverage among ASlncRNAs against the average coverage in 1000 random samples of noASlncRNAs. Significantly enriched TE classes *(with adjusted p-value <= 0.01)* are highlighted in green whereas the depleted ones are highlighted in red.

2.4. Conclusions

In conclusion, the modular pipeline proved extremely useful in the identification, classification and generation of valuable resource of data for the study of ASlncRNAs. The repeat content analysis revealed the fraction of genes containing TEs differed largely among different species, where human and mouse contained the highest fraction of genes with SINE elements. Further, TEs coverage enrichment analysis revealed ASlncRNAs are significantly enriched for SINE derived sequences in contrast to noASlncRNAs. This is an intriguing observation specially because the SINE repeats are identified to be the effector domain in *AS-Uchl1* (Carrieri et al., 2012) and synthetic SINEUPs (Zucchelli et al., 2015a)..

Considering the significant enrichment of SINE repeat coverage specifically among ASlncRNAs in human and mouse, I decided to further focus only on human and mouse for detailed exploration of SINE family and subfamily specific coverage enrichment analysis, with the aim to identify the contributions of individual SINE elements in the sequence composition of ASlncRNAs and ultimately their functional associations within the spectrum of human and mouse lineage. From here on, I have not considered to further study any of the non-mammalian species as they do not show similar SINE specific coverage enrichment and have generally lower percentage of exonized TEs which could be due to the relative lower number of available annotated ASlncRNAs in comparison to human and mouse.

Chapter 3

Analysis of SINE coverage enrichment among ASlncRNAs in human and mouse with respect to noASlncRNAs

3.1. Introduction

SINEs are the widespread TEs among eukaryotic organisms. They can be found in most of the flowering plants as well as mammals, reptiles, fishes and in many invertebrates such as cephalopods, sea squirts, sea urchins, nematodes and certain insects. However, the genomes of *Drosophila* and many unicellular eukaryotes do not contain SINE repeats (Kramerov & Vassetzky, 2011). Unlike other TEs that are transcribed by RNA polymerase II, SINEs are transcribed by RNA polymerase III (pol III) and contain a pol III promoter in their sequence. SINEs are non-autonomous retrotransposons and rely on LINE reverse transcriptase for retrotransposition (Dewannieux, Esnault, & Heidmann, 2003). The structure of most the SINEs contain three modules- 5' head, body and 3' tail. The head of SINE elements are known to be derived from tRNA, 7SL RNA or 5S rRNAs, whereas the tail is a sequence of variable lengths consisting of simple repeats (*reviewed in detail by* Kramerov & Vassetzky, 2011).

Higher mammals such as human and mouse are known to have abundance of SINE elements in comparison to non-mammalian vertebrates and other invertebrates (Sela et al., 2010). With the split of primate-rodent lineages, mouse has accumulated diverse SINE type sequence in comparison to human genome that is colonized by a smaller number of SINE types. For example, apart from the common ancestral SINE MIRs *(a tRNA-derived Mammalian-wide Interspersed Repeats)* (Silva et al., 2003; Smit & Riggs, 1995), the mouse lineage is exposed

to four major distinct SINE families: SINE B1, B2, ID and B4. Whereas, in case of human, SINE *Alu* is the only known major SINE family apart from the ancestral MIR (Mouse Genome Sequencing Consortium, 2002). Mouse SINE B1 and human SINE *Alu* repeats are known to have originated from a common source, 7SL RNA. The 7SL RNA is an abundant cytoplasmic RNA with a known function in protein secretion as a component of signal recognition particle (SRP) that recognizes and targets the specific proteins to endoplasmic reticulum in eukaryotes, and to plasma membrane in prokaryotes.(Walter & Blobel, 1982). Oher mouse specific SINE families such as, SINE B2 are known to have derived from t-RNAs, SINE ID are known to be derived from a neuronally expressed RNA gene called *BC1* and lastly, SINE B4 family are suggested to resemble the fusion of B1 and IDs as most of the ID repeats are found within 50 bp distance from the B1 repeats (Mouse Genome Sequencing Consortium, 2002). SINE retrotransposons are known to have played an important role in the early evolution and reshaping of the genomes of human and mouse lineages *(discussed in section 1.6)*. In order to understand the contributions of SINE elements to the transcriptomes of human and mouse, it is important to analyze their sequence coverage, which would also reveal their contributions to the sequence composition of different ASlncRNA and noASlncRNA transcript groups.

In the previous chapter, we have seen ASlncRNA transcripts are significantly enriched for SINE elements, whereas other TEs show relatively lower coverage among ASlncRNA in comparison to noASlncRNA *(Figure 2.4 a,b)*. This suggests that the ASlncRNA transcript sequences tend to retain SINE repeats and hence might be functionally associated. In this chapter, I have described subsequent analysis I performed in order to identify if a specific family/subfamily of SINE elements contribute significantly to the ASlncRNA transcript sequences. I further compared the positional distribution of SINE elements within the

ASlncRNA and noASlncRNA transcripts and identified the portion of the sequence for each SINE element that is in frequent overlap with ASlncRNA and noASlncRNAs.

3.2. Materials and Methods

3.2.1. SINE family and subfamily coverage enrichment analysis

In order to infer the contributions of each SINE family in previously observed cumulative coverage of SINE repeat class shown in *(Figure 2.4 a,b)*, their coverage percentages are computed by quantifying the total number of nucleotides covered by each SINE family within the total number of SINE covered nucleotides in different transcript groups. The computed percentages are then represented into charts for further interpretations. This analysis is performed using the TEs coverage analysis module of the pipeline which can also be used to analyze the coverage of multiple TE classes simultaneously by passing the names of the TEs class as a list argument in R, for example, *'c("SINE", "LINE", "LTR")'* to analyze the coverage for SINE, LINE and LTR families together.

For computing the statistical enrichment of the observed coverage for each SINE family and its corresponding subfamilies among ASlncRNA with respect to noASlncRNAs, a randomization analysis is performed, wherein 1000 random samples of noASlncRNA transcripts are generated *(sample size, n = total transcripts in ASlncRNA group).* For each of the generated random samples, the coverage of SINE family/subfamily is computed considering the number of nucleotides covered by each SINE element, out of the total number of nucleotides in each of transcript groups. The *mean* and *standard-deviation (sd)* of the accounted coverage for each SINE element in 1000 samples is then compared against their actual coverage in ASlncRNAs to generate Z-score as explained previously in section 2.2.4.3,

using the formula: $\mathbf{z = (X - \mu) / \sigma}$, where, *z* is the Z-score, *X* is the SINE coverage observed among ASlncRNAs, μ is the mean of SINE coverage in 1000 samples of noASlncRNAs and finally σ is the observed standard deviation in SINE coverage for the population of noASlncRNA random samples. The Z-score thus obtained is used to calculate p-values using the *pnorm()* function in R (once again, as explained previously in section 2.2.4.3) considering the two sided test like this: **p-value = 2 * pnorm(-abs(z))**. The generated p-values are then adjusted using the FDR method. Further, the fold-change difference between the coverage observed in random sample and ASlncRNAs are calculated using the *foldchange()* in R. The foldchange values are then converted to log2 ratio values using the *foldchange2logratio()* function in R. I have used thus generated log2 ratio values for the graphical representation of the observed nucleotide level coverage enrichment for all SINE family/subfamilles *(Figure 3.2)*.

3.2.2. Identification of SINE covered regions across the ASlncRNAs and noASlncRNAs

In order to compare the positional distribution of SINE covered regions within the ASlncRNA and noASlncRNA transcripts, their genomic coordinates are normalized to a common scale ranging from *0 to 1*, using the transcript start and end coordinates

- normalized repeat start = (repeat start – transcript start)/(transcript end – transcript start)
- normalized repeat end = (repeat end – transcript start)/(transcript end – transcript start)

Next, the number of transcripts containing repeats with common normalized range of start and end positions are identified making use of the *hist()* function in R (The R Development Core Team, 2004). Results are then represented into charts for interpretation, as percentage of transcripts that contain repeats at specific positions ranging from *0* to *1* scale, where *0*

represents the transcript start and *1* represents the transcript end positions respectively for ASlncRNAs and noASlncRNAs.

3.2.3. Identification of SINE regions under frequent overlap with ASlncRNAs and noASlncRNAs

Given that the SINE elements are significantly enriched among ASlncRNAs with respect to noASlncRNAs it would be further interesting to investigate if the SINE elements overlapping to ASlncRNAs are more complete *(full length)* or if specific region of the SINEs elements preferentially embed among ASlncRNAs with respect to noASlncRNAs. In order to infer this, the total number of repeats with common overlapping start and end coordinates *(with respect to transcripts)* among ASlncRNAs and noASlncRNAs are accounted, using the *hist()* function in R. The number of repeats overlapping at specific portion across its length, against the AslncRNAs and noASlncRNAs are then represented into charts as shown in figure 3.4.

3.3. Results and discussions

3.3.1. SINE family coverage

The coverage of individual SINE families across different transcript groups for human and mouse are shown in figure 3.1 *a* and *b* respectively. From the charts, it is clearly observable that *Alu* and MIR repeats are the two major SINE families that are predominantly present in all transcript groups in human. Also *Alus* particularly show a higher percentage of coverage among ASlncRNAs. Whereas in case of mouse, B1 *(annotated as Alu by RepeatMasker)*, B4, B2 and MIR are the four major families present across all transcript groups, and the coverage percentages of B1, B2 and B4 families are particularly higher among ASlncRNAs *(Figure 3.1 b).*

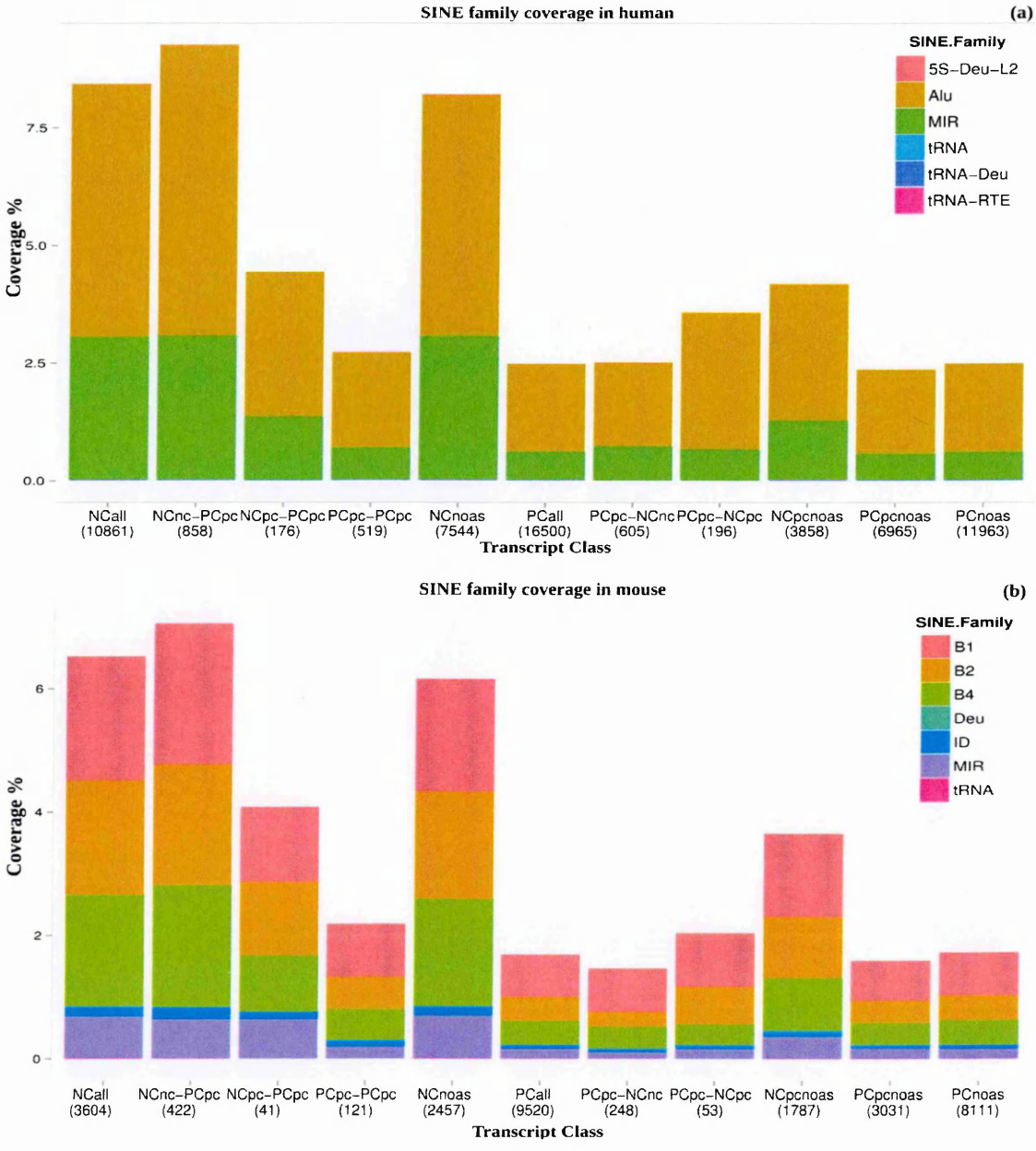


Figure 3.1 | SINE family coverage. The above charts represents the percentage of total transcript nucleotides represented by the different TE classes *(y-axis)* among the different transcript categories. The total number of transcripts in each category is indicated below the category name within brackets *(x-axis)*. Charts are for (a) human, (b) mouse respectively.

3.3.2. SINE family coverage enrichment analysis

The charts shown in figure 3.1 shows a comparable difference in the percentage of coverage for SINE families between ASlncRNAs and noASlncRNAs. However, in order to know if this difference in coverage is not just by chance, there is a need of a statistical validation. As a consequence, I performed the coverage enrichment analysis discussed in methods 3.2.1. Interestingly, the results of the coverage enrichment analysis revealed that the *Alu* and *MIR* repeats are indeed significantly enriched among ASlncRNAs in comparison to noASlncRNAs *(Table 3.1)*. Although, this is clearly observable in case of *Alu (figure 3.1 a)*, MIR SINE family do not show a big difference in coverage between *NCnc-PCpc* and *NCnoas* group of transcripts. However, keeping into account the fact that the ASlncRNAs used in the coverage enrichment analysis are made by the union of *NCnc-PCpc* and *NCpc-PCpc* transcript groups and noASlncRNAs are the union of *NCnoas* and *Ncpcnoas* transcripts, I could deduce that the MIR enrichment among ASlncRNAs is mainly contributed by the coverage difference between *NCpc-PCpc* and *NCpcnoas* group of transcripts belonging to ASlncRNAs and noASlncRNAs respectively. Other interesting known fact associated to the ASlncRNA enriched MIRs repeats is that, they predominantly carry TSS (transcription start sites) for *cis natural antisense* transcripts in human (Conley, Miller, & Jordan, 2008) and are also significantly enriched for transcription factor binding sites (Polavarapu et al., 2008). Given that, the MIRs are the ancestral TEs, selection might have acted to conserve their function in human lineage.

On the other hand, in case of mouse the coverage enrichment analysis revealed SINE B1, B2 and B4 SINE families are significantly enriched among ASlncRNA with respect to noASlncRNAs *(Table 3.1)*.

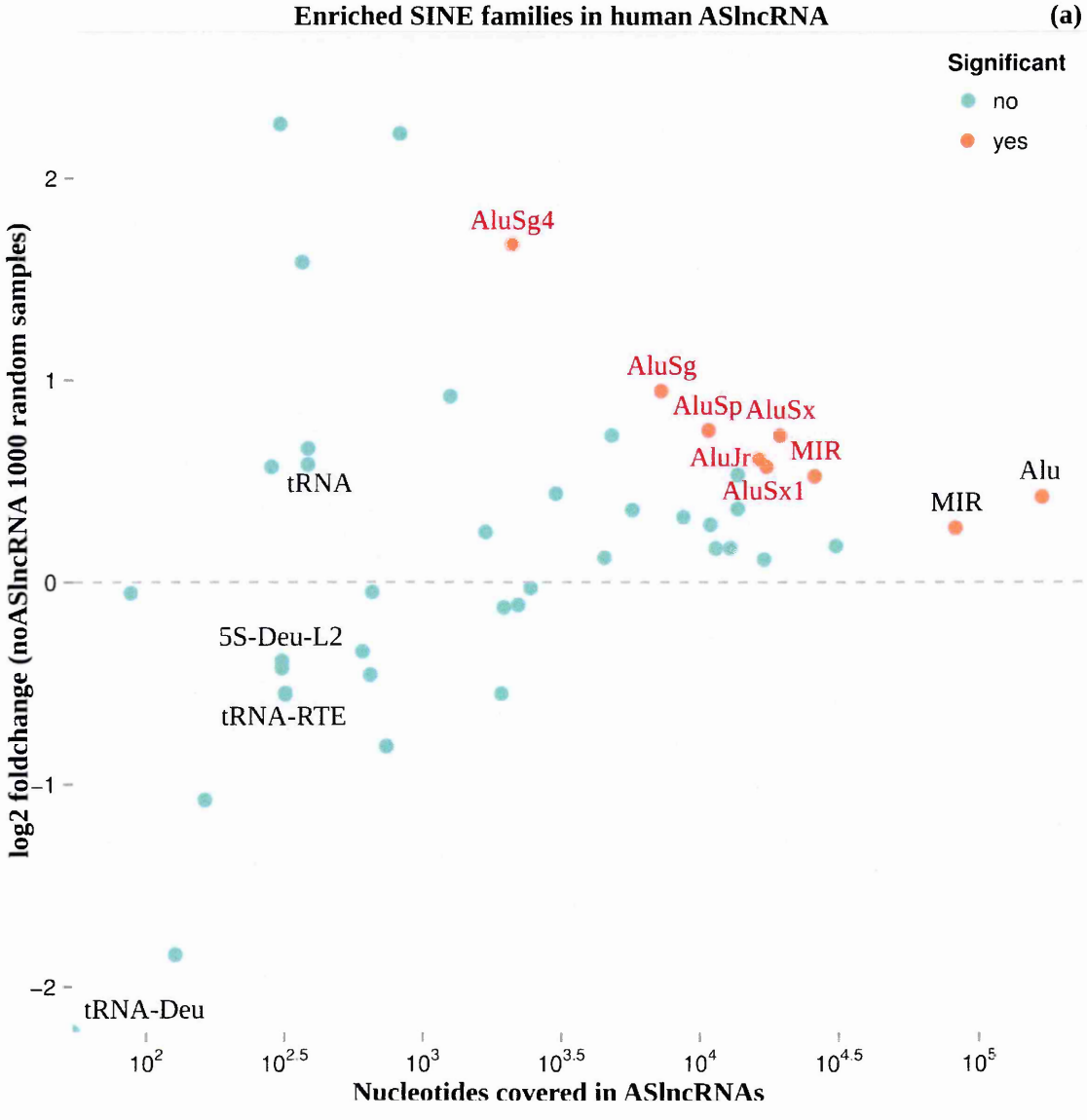
Species	SINE Family	ASlncRNA (Coverage %)	noASlncRNA (Coverage %)	Z-score	p-value	p-value (Adjusted)
Human	5S-Deu-L2	0.0100778265	0.0132001166	-0.3744169528	7.08e-001	8.06e-001
	Alu	5.5161609146	4.1173002854	6.0489327212	1.46-009	4.81e-008
	MIR	2.7002976252	2.240781752	3.8478252197	1.19e-004	1.97e-003
	tRNA	0.0124490798	0.0083210616	0.6947908745	4.87e-001	7.66e-001
	tRNA-Deu	0	0.0028729978	-0.7927265142	4.28e-001	7.43e-001
	tRNA-RTE	0.0103742332	0.0151683402	-0.5408141947	5.89e-001	8.02e-001
Mouse	B1	2.1746722331	1.59995614	3.4805421712	5.00e-004	7.51e-003
	B2	1.8752648205	1.3683923554	3.3638121206	7.69e-004	7.69e-003
	B4	1.8551781486	1.2952707572	4.0892525601	4.33e-005	1.30e-003
	Deu	0	0.0046250848	-0.4768270245	6.33e-001	8.34e-001
	ID	0.1921970912	0.1242190379	2.4971873343	1.25e-002	9.39e-002
	MIR	0.6355081073	0.5162556696	1.6488373366	9.92e-002	4.25e-001
	tRNA	0.0070211771	0.0110234087	-0.3858623431	7.00e-001	8.34e-001

Table 3.1 | SINE family coverage enrichment. Above table contains the coverage, Z-scores and p-values generated from the enrichment analysis about the coverage of SINE families among ASlncRNAs against the mean coverage in 1000 random samples of noASlncRNAs. Significantly enriched TE classes *(adjusted p-value <= 0.01)* are highlighted in green for each species.

3.3.3. Coverage enrichment analysis for SINE subfamilies

Till now we saw the differences in the coverage of SINE families among ASlncRNA and noASlncRNAs and identified specific SINE families that are significantly enriched among ASlncRNA sequences in human and mouse. However, as we previously discussed, with the split of primate-rodent lineages, SINE elements have undergone a species specific evolution. Hence, both mouse and human genome have accumulated diverse SINE subfamilies/elements within each SINE family. AS a consequence, I was next interested to identify the contributions of individual subfamilies/elements in cumulative coverage of SINE family among ASlncRNAs, because this would not only filter out important individual SINE subfamily/element contributing to ASlncRNA, but also shed light on the dynamics of SINE elements in human and mouse lineages. For this, I once again performed the coverage enrichment analysis, but this time considering each SINE subfamilies/elements as discussed in methods *3.2.1*. The results of this analysis revealed that in case of human, there are eight *Alu* elements that are significantly enriched among ASlncRNA out of 50 annotated *Alus* as per RepeatMasker *(Figure 3.2 a)*. The enriched *Alu* elements belong to the two major *Alu* subfamilies, Alu-J and Alu-S (Jurka & Smith, 1988). Alu-J is the oldest dimeric subfamily and contains *AluJr* element. Whereas, Alu-S is relatively younger dimeric SINE subfamily that differ from Alu-J in several region of their sequences. Alu-S comprises *AluSx, AluSx1, AluSp, AluSg* elements that are identified to be significantly enriched among ASlncRNAs *(Table 3.2)*. Intriguingly, *AluSx* elements are also known to be enriched for several RNA binding proteins (RBPs) (Kelley, Hendrickson, Tenen, & Rinn, 2014).. Lastly, Alu-Y is the youngest known subfamily containing *AluYh3* and *AluYa8 (Table 3.2)* enriched elements. I chose not to highlight *AluYh3* and *AluYa8* elements in *figure 3.2a* because, although they are significantly enriched among ASlncRNAs in contrast to noASlncRNA, they cover less than 900 nt in total

(Table 3.2). Similarly, out of five annotated MIR subfamilies, *MIR* is identified as the significantly enriched among ASlncRNA sequences other four being *MIR1_Amn*, *MIR3*, *MIRb*, and *MIRc (Table 3.2; Figure 3.2a)*. On the other hand, in case of mouse the coverage enrichment analysis revealed, *ID_B1, B3A* elements from B4 and B2 SINE family respectively and *B1_Mus1* and *PB1D10* elements from B1 SINE family to be significantly enriched among ASlncRNAs *(Figure 3.2b, Table 3.2)*



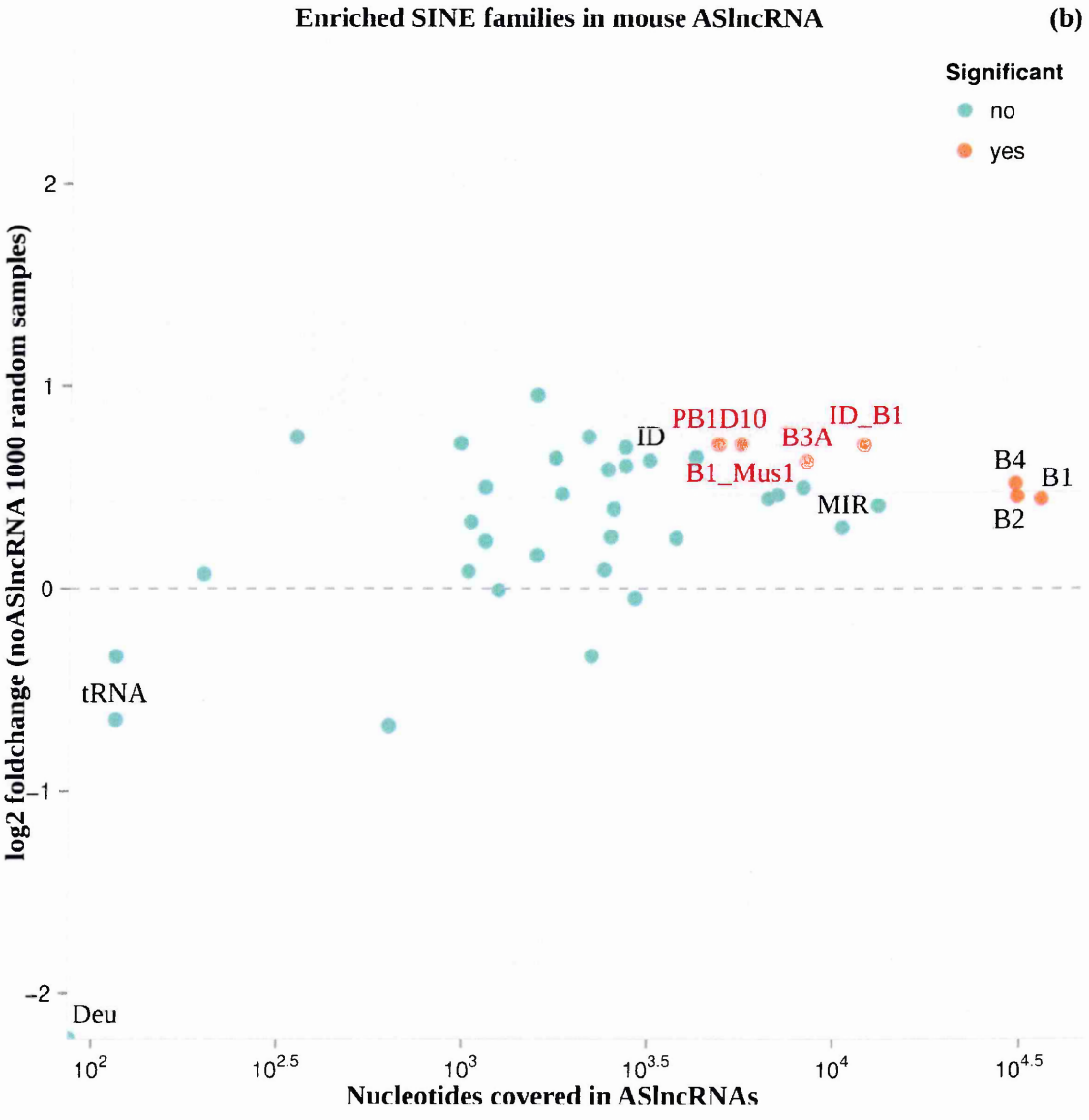


Figure 3.2 | SINE family coverage enrichment. In above charts for (a) human and (b) mouse, *y-axis* represents the fold-change difference between the observed coverage for SINE families and subfamilies annotated by RepeatMasker, among ASlncRNAs and 1000 random samples from noASlncRNAs. Whereas, *x-axis* represents the number of nucleotides covered by each SINE family and subfamilies *(at least >=2000 nt)* in ASlncRNAs. The dots represent individual SINE family or

subfamilies. They are colored in red only if the corresponding SINE family or subfamily is significantly enriched *(adjusted p-value <= 0.08)* among ASlncRNAs, or else are shown in green. All SINE families are labeled in black whereas only the significantly enriched SINE subfamilies are labeled in red.

Species	SINE Subfamily	SINE Family	Nt. covered	ASlncRNA (Coverage %)	Z-score	Fold-change (log2)	p-value (Adjusted)
Human	MIR	MIR	25662	0.8451541957	4.3879997496	0.5243268861	3.66e-004
	AluSx	Alu	19262	0.6343761249	4.3340325468	0.7256651849	4.39e-004
	AluSx1	Alu	17221	0.5671576808	3.1472700134	0.5724657386	3.10e-002
	AluJr	Alu	16151	0.5319182221	3.5182751985	0.6099725191	1.02e-002
	AluSp	Alu	10642	0.3504844108	3.2750071817	0.751775682	2.11e-002
	AluSg	Alu	7162	0.2358738348	3.5479382915	0.9460587234	9.31e-003
	AluSg4	Alu	2065	0.0680088619	3.2976507625	1.6728910028	1.99e-002
	AluYh3	Alu	806	0.0265448633	3.6876666853	2.219613162	5.57e-003
	AluYa8	Alu	79	0.0026017918	5.4653081587	4.3878461738	1.85e-006
Mouse	ID_B1	B4	11909	0.7270886781	3.6976091845	0.7088470809	3.85e-003
	B3A	B2	8370	0.5110195848	2.7443967427	0.6263520127	7.58e-002
	B1_Mus1	B1	5596	0.341656582	2.80668967	0.7111096672	6.46e-002
	PB1D10	B1	4867	0.2971484252	3.1233719934	0.7120631597	2.85e-002

Table 3.2 | SINE subfamily coverage enrichment. Above table contains the information about total number of nucleotides covered, coverages, Z-scores, fold-change differences in coverage and p-values for all SINE subfamilies that are significantly enriched among ASlncRNAs *(adjusted p-value <=0.08)* in human and mouse *(data sorted by decreasing coverage %).*

The outcome of the SINE subfamily/element specific coverage enrichment is very interesting particularly because both *Alu* and B1 SINE families are identified to be significantly enriched among ASlncRNA sequences in human and mouse respectively. As we previously discussed *Alus* and B1 elements are originated from a common source of origin called as 7SLRNA, before the divergence of primate-rodent lineages and have followed different evolutionary routes since the split. Additionally, the significantly enriched *PB1D10* elements belonging to B1 SINE family in mouse are the progenitor for the first modern B1 SINE elements, hence were also referred as proto B1 elements (Quentin, 1994; Veniaminova, Vassetzky, & Kramerov, 2007). Taken together, these observations suggest that the ASlncRNA sequences in human and mouse have evolved similarly and are enriched for similar SINE elements. Other interesting facts associated with mouse B1 SINE family includes that the B1 elements are also known to show different structural features within different rodent families and *B1_Mus* subfamily is a mouse specific B1 element. Additionally, B1 elements are also known to form dimeric SINEs along with ID elements, where *ID_B1* subfamily in mouse represent such dimers (Veniaminova et al., 2007). Lastly, the enrichment of *B3A* of SINE B2 family among ASlncRNAs is also an intriguing observation, specially because the SINE B2 element is reported as the effector domain in *AS-Uchl1* by *Carrieri et al., 2012,* that belongs to *B3* subclass and *B3A* element is very similar to B3 (Repbase Update – GIRI). Hence SINE B3A elements of SINE B2 family in mouse could represent the potential candidates for the effector domain similar to that of SINE B2 element embedded in *AS-Uchl1*. However, this requires further investigation in terms of their role among the various ASlncRNAs that are similar to the modular *AS-Uchl1* and synthetic SINEUPs (Carrieri et al., 2012; Zucchelli et al., 2015a).

3.3.4. Positional distribution of SINE elements within ASlncRNAs and noASlncRNAs

We just identified that the SINE elements are enriched among ASlncRNA sequences in human and mouse also share the common origin. Additionally, the significantly enriched B3A element of SINE B2 family in mouse are very similar to the SINEB2 elements that is reported as the effector domain in *AS-Uchl1*. This suggests that the enriched SINE elements in mouse and human might potentially act as the effector domains among ASlncRNAs. However, if this is the case then the ASlncRNAs should also resemble to *AS-Uchl1* in terms of the location of SINE element with in the transcript body, which is near to the 3' end for *AS-Uchl1*. As a consequence, aiming to infer the positional distribution of SINE elements within the transcript body of ASlncRNAs with respect to noASlncRNA, I used the previously generated *tx.level* and *repeat.level* tables to access genomic coordinates of transcripts and their overlapping repeat elements. The positional mapping of SINE overlap regions within ASlncRNAs and noASlncRNAs are performed as explained in section ***3.2.2***.

The percentage of total transcripts in ASlncRNA and noASlncRNAs containing SINE elements at specific position across their transcript lengths are represented in *figure 3.3*. As expected, the ASlncRNA transcripts in both human and mouse show a noticeable difference in the peaks for SINE repeat positional distribution against noASlncRNAs *(Figure 3.3 a,b)*. The noASlncRNA transcripts contain SINE repeats throughout the transcript body with prominently higher percentage of transcripts containing SINE elements specifically at the 5' and 3' ends, whereas a higher fraction of ASlncRNAs tend to have SINE overlaps particularly near the 3' ends of the transcripts *(the peaks distribution of individual SINE families among ASlncRNAs are shown in figure 3.3 c and d for human and mouse respectively)*. This is also an expected characteristic of ASlncRNAs because they would require their 5' end sequence to be

uninterrupted by repeat insertions, so that they could recognize and overlap to the specific target sense mRNAs to form S/AS pair of transcripts. In sum, the analysis of the positional distribution of SINE elements within ASlncRNA transcripts of human and mouse are very similar to the modular *AS-Uchl1* (Carrieri et al., 2012).

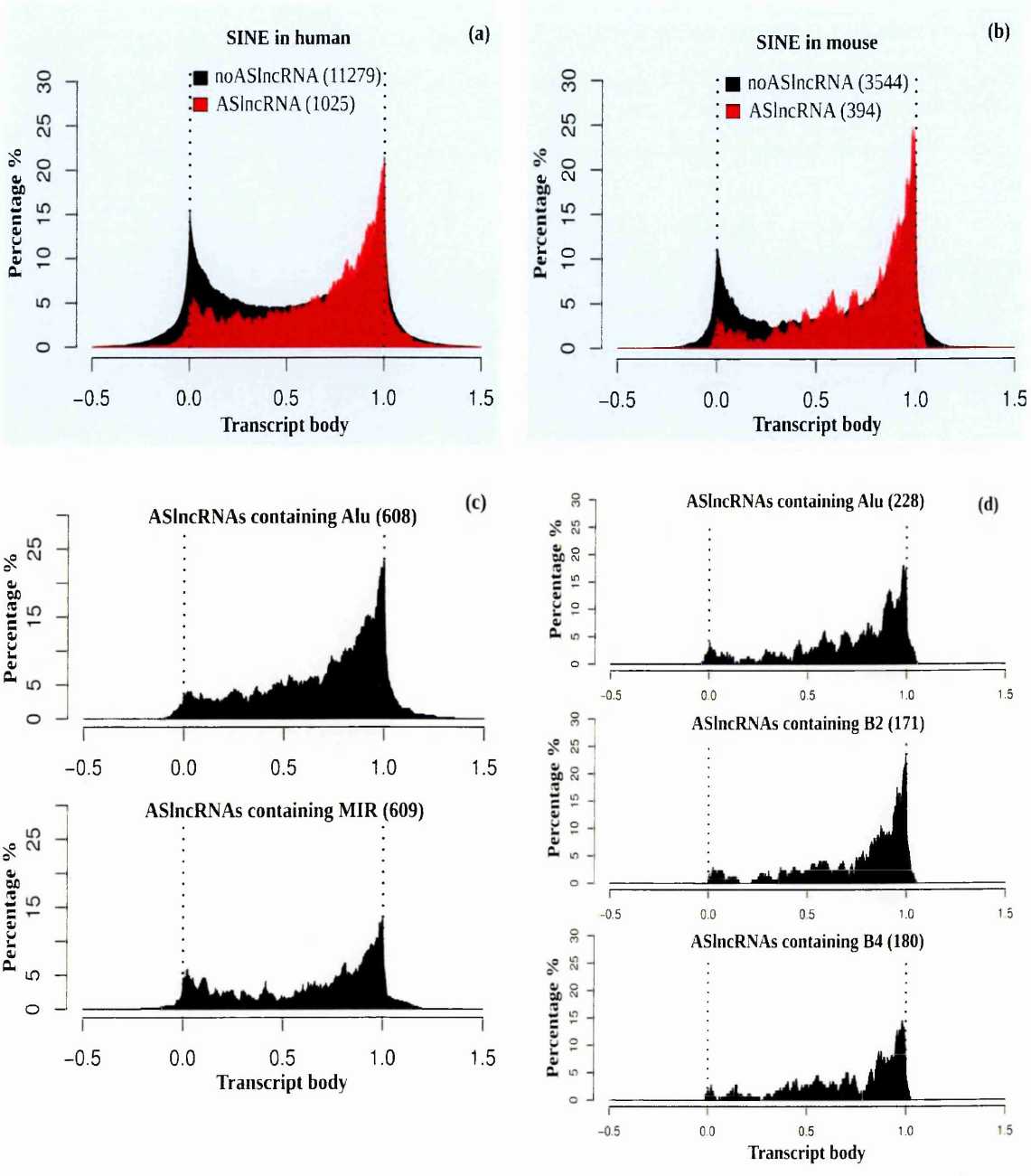


Figure 3.3 | SINE coverage peaks across the transcripts. Above figures *(a), (c)* and corresponds to human, whereas figures (*b) and (d)* corresponds to mouse. *(a)* and *(b)* represents the percentage of ASlncRNA *(in red)* and noASlncRNA *(in*

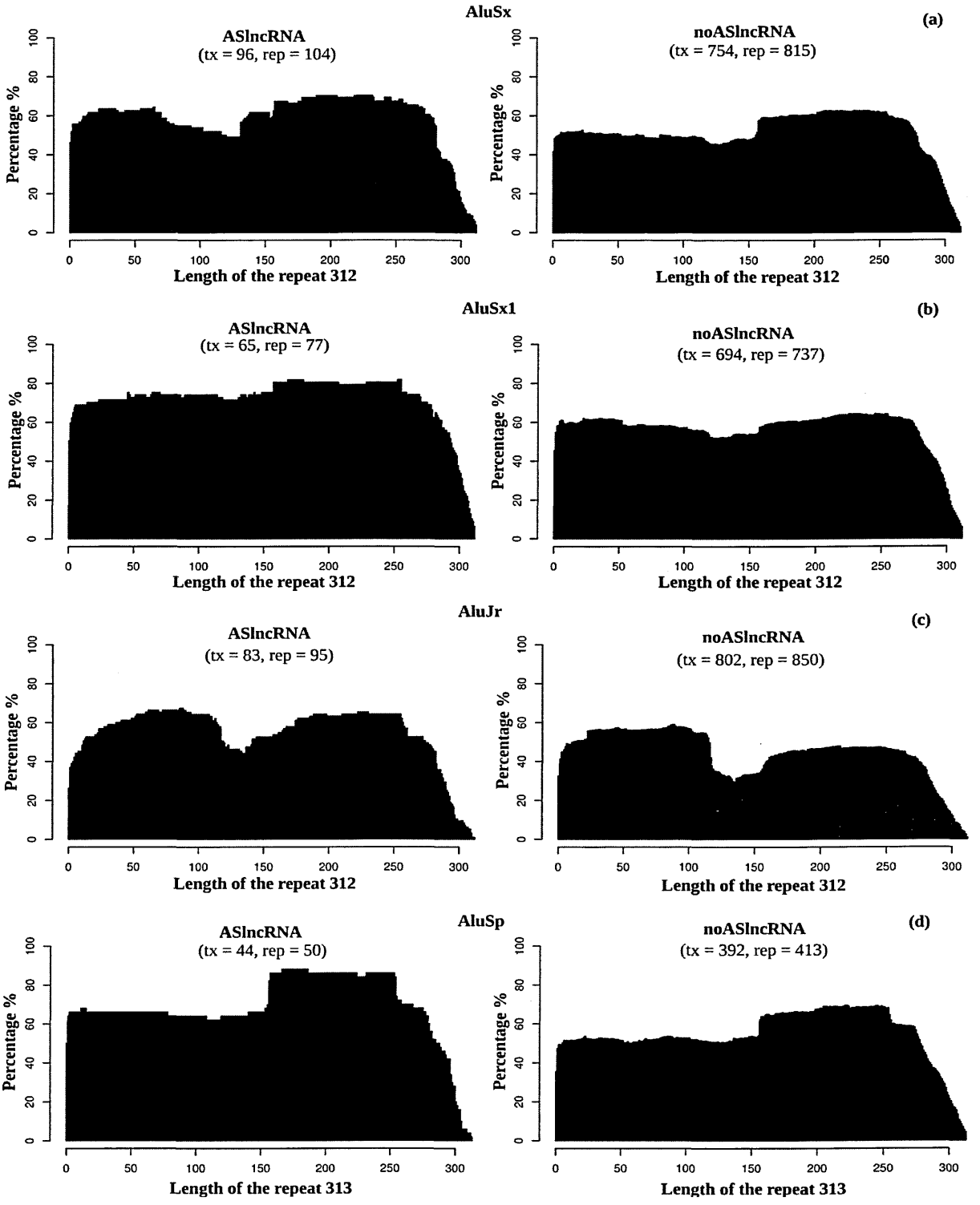
black) in *y-axis* (total number of transcripts in each group are shown in the legends) overlapping to SINE family of repeats *("Alu", "MIR" in case of human and "Alu", "B2" and "B4" in case of mouse)* at specific regions within the transcript body shown in *x-axis*. Here, *0 and 1* denotes the transcript start and end points respectively, whereas the scale ranging between *'-0.5 - 0'* and *'1 - 1.5'* represent the flanking regions near to *5'* and *3'* ends of the transcripts. Rest of the figures below are similar representation for ASlncRNAs overlapping to individual SINE families for human and mouse respectively.

3.3.5. Region of SINE elements under frequent overlap with ASlncRNAs and noASlncRNAs

So far we analyzed the coverage enrichment of SINE families and inferred their positional distribution within the ASlncRNAs and noASlncRNAs and identified that the ASlncRNAs are enriched for specific SINE subfamilies/elements in human and mouse, also share a common origin and resemble to the modular *AS-Uchl1* in terms of the positional distribution of SINE element within the transcript body. Previously, we also identified that ASlncRNA tend to retain SINE elements with higher percentage of SINE coverage while other TEs show lower coverage percentage in comparison to noASlncRNAs. Based on these observation one could hypothesize that the SINEs are in positive selection when they are invading to the ASlncRNA sequences, whereas SINEs invading to noASlncRNAs are under no selective pressure with mutations mangling the sequences of each insertion thereby resulting in related elements that are of different length, incomplete structure or sequence. As TEs are in general known to be under no selective pressure and hence prone for mutations leaving behind the sequence which

is often beyond recognition by sequence similarity techniques (Kaminker et al., 2002; Turlan, Loot, & Chandler, 2004; Van De Lagemaat et al., 2005).

if the above hypothesis holds true, then the SINE elements embedded to ASlncRNAs are more likely to have complete sequence, whereas the SINEs among noASlncRNAs would be highly mutated and have incomplete sequence. To check this, I decided to compare the SINE sequence overlap frequency between the ASlncRNAs and noASlncRNAs as discussed in methods **3.2.3.** And represented the percentage of SINE elements overlapping at specific region across its length, against ASlncRNA and noASlncRNA in the figure *3.4* and *3.5*, for human and mouse respectively. In the charts, it is clearly observable that the SINE elements overlapping to ASlncRNAs and noASlncRNAs do not show any difference in the overlap region, as the sequences of SINE elements in ASlncRNA and noASlncRNAs appears to be equally mutated with similar incomplete sequences. This shows that even though ASlncRNAs are significantly enriched for specific SINE elements in comparison to noASlncRNAs, SINE elements embedded to ASlncRNAs are are not under positive selection.


AluSx
(a)
ASlncRNA
(tx = 96, rep = 104)
noASlncRNA
(tx = 754, rep = 815)
AluSx1
(b)
ASlncRNA
(tx = 65, rep = 77)
noASlncRNA
(tx = 694, rep = 737)
AluJr
(c)
ASlncRNA
(tx = 83, rep = 95)
noASlncRNA
(tx = 802, rep = 850)
AluSp
(d)
ASlncRNA
(tx = 44, rep = 50)
noASlncRNA
(tx = 392, rep = 413)
Percentage %
Length of the repeat 312
Length of the repeat 313

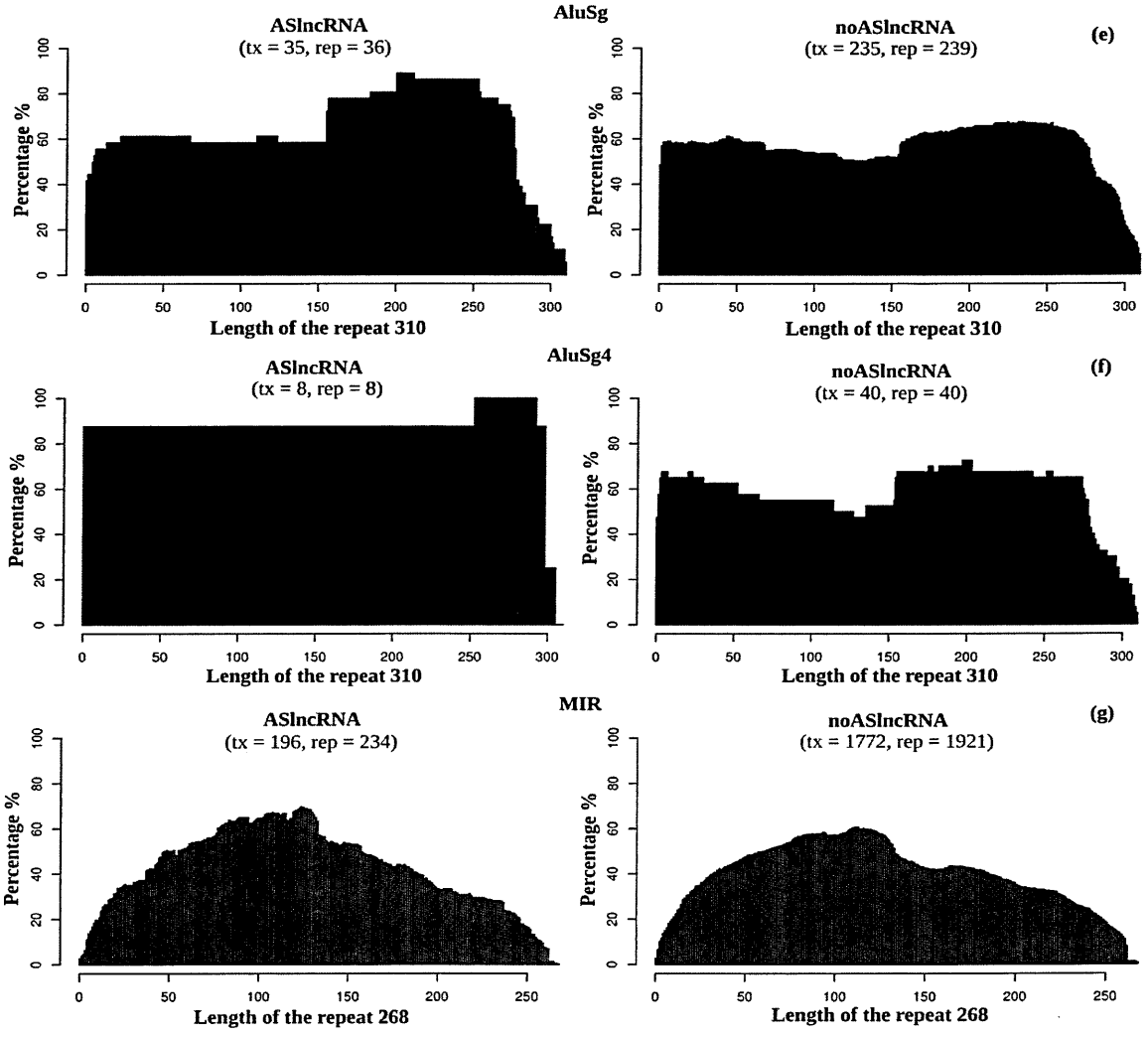


Figure 3.4 | Region of SINE elements under overlap with ASlncRNAs and noASlncRNAs in human. In the above charts *y-axis* represents the percentage of total repeats *(shown in the brackets below the transcript category name)* having an overlap at specific region across its reference length. The reference length is indicated on the *x-axis* among ASlncRNAs *(left)* and noASlncRNAs *(right) for each* for each SINE subfamilies *(subfamily names are shown on the top of each pair of charts).*

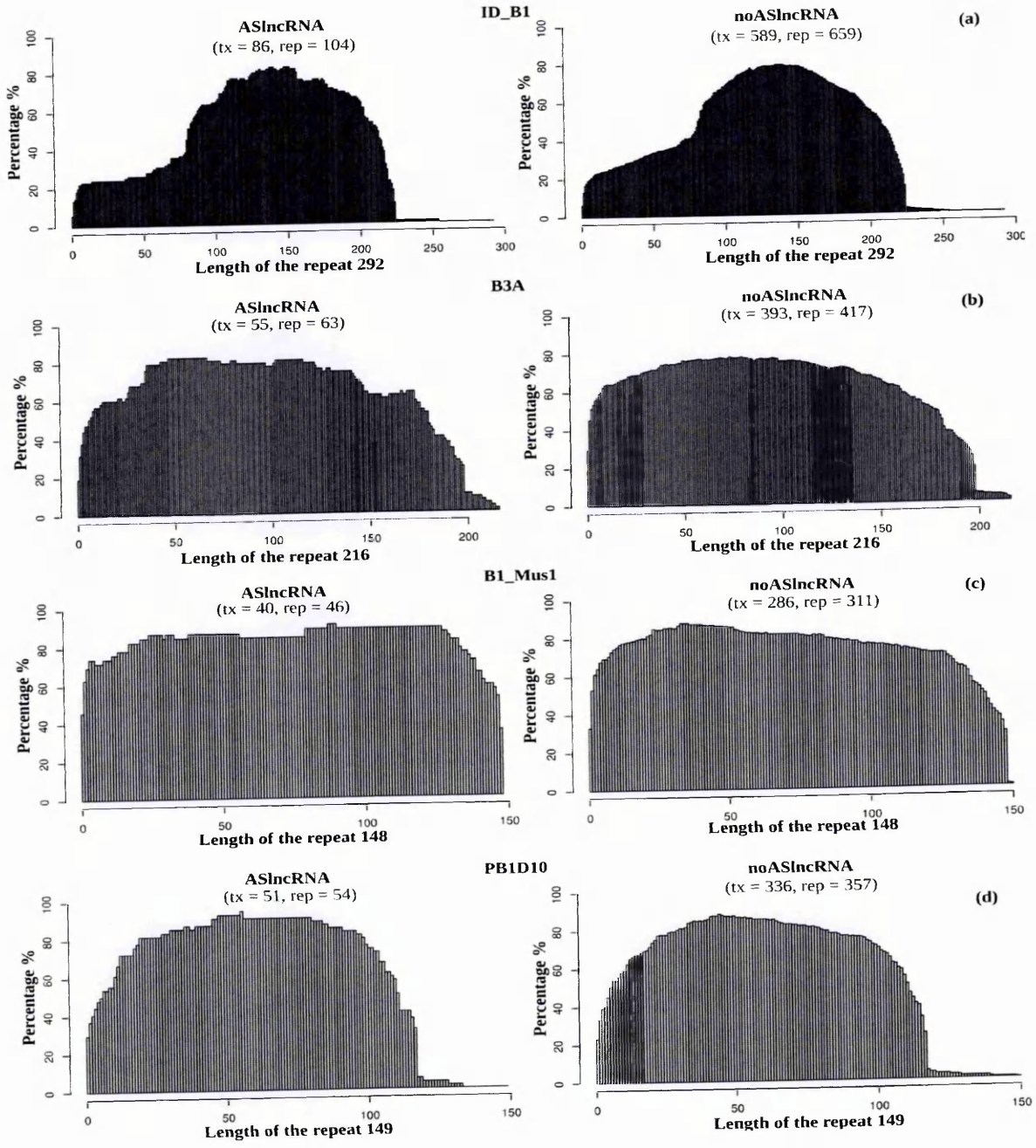


Figure 3.5 | Region of SINE elements under overlap with ASlncRNAs and noASlncRNAs in mouse. *(Figure description similar as that of figure 3.4).*

3.4. Conclusions

Based on the list of analysis described in this chapter the main conclusion which can be drawn is that the ASlncRNAs are significantly enriched for the coverage of specific SINE families that includes *Alu* and MIR in human, and B1, B2 and B4 in mouse. Looking into the coverage of each SINE subfamily/element revealed majority of the enriched elements in human and mouse are the oldest among all subfamilies and are derived from a common origin (7SL RNA) in both human and mouse. This suggests that the ASlncRNAs in human and mouse have undergone similar dynamics of evolution, where the insertions of similar SINE element have significantly contributed to the ASlncRNA sequences in contrast to noASlncRNAs. Additionally, SINE B3A element is identified as significantly enriched SINE B2 subfamily contributing to the ASlncRNA sequences in mouse, which is very similar to SINE B2 element that is identified to be the effector domain in *AS-Uchl1*. This also suggests that the ASlncRNAs contaiing SINE B3A elements could potentially be the good candidates to be tested for *AS-Uchl1* like activity ASlncRNAs in human and mouse also show a similar trend of SINE distribution across their lengths which also closely resembles to the modular architecture described for functional *AS-Uchl1*. Finally, the comparative analysis of SINE overlap region within ASlncRNAs and noASlncRNAs showed that even though ASlncRNAs are enriched for specific SINE subfamily/.elements and show similar distribution of SINE elements across their transcript lengths as that of modular *AS-Uchl1*, the SINEs embedded to ASlncRNAs are not under positive selection.

Chapter 4

Analysis of the modular nature of ASlncRNAs

4.1. Introduction

The *AS-Uchl1* is the first reported ASlncRNA which is shown to exert a post-transcriptional protein up-regulatory activity over its sense overlapping *Uchl1* mRNA during cellular stress condition *(Figure 1.2)* (Carrieri et al., 2012). Its activity mainly relies on its two domains – 1) the 5' binding domain and 2) the 3' effector domain. Both the domains have their own important characteristics which are required in order to observe the functional activity of *AS-Uchl1*, the deletion of either of them resulted in the loss of UCHL1 protein upregulation *(Figure 4.1 a). AS-Uchl1* is therefore proposed to represent a new functional class of natural modular ASlncRNA that can activate translation of sense overlapping transcripts. They are also referred as SINEUPs, because they require inverted SINEB2 sequence to upregulate translation in a gene-specific manner (Zucchelli et al., 2015a).

One of the important characteristics of modular *AS-Uchl1* is the presence of an inverted SINEB2 element near the 3' end, which acts as an effector domain. The specific "inverted" orientation of SINEB2 with respect to *AS-Uchl1* is an important required characteristic, as the mutant construct of *AS-Uchl1* with flipped SINEB2 *(SF in Figure 4.1b)* failed to function. In case of the 5' binding domain, a minimal overlapping sequence is required for targeting *Uchl1* mRNA. A synthetic construct containing only 73 *nt* sequence of the 5' binding domain, adjacent to the inverted SINEB2 element was identified to be still able to upregulate UCHL1 protein levels *(Figure 4.1c).*

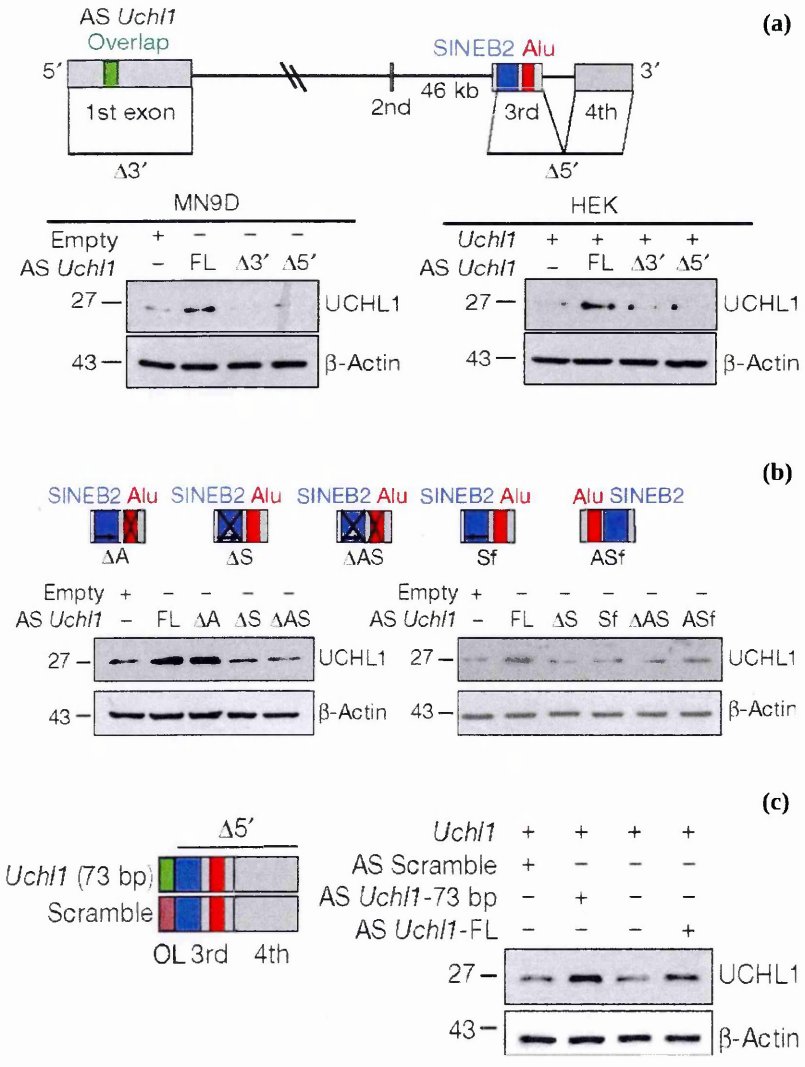


Figure 4.1 | Characteristics of the modular *AS-Uchl1*. (a) Full-length (FL) antisense Uchl1 is required for regulating endogenous (MN9D cells, left panel) and over-expressed (HEK cells, right panel) UCHL1 protein levels. Scheme of Δ5′ or Δ3′ deletion mutants is shown and the overlap region is indicated in green (b) Inverted SINEB2 is sufficient to control endogenous UCHL1 protein levels in MN9D cells. Scheme of mutants is shown in 5′ to 3′ orientation. ΔA, ΔAlu; ΔS, ΔSINEB2; ΔAS, ΔAlu+SINEB2; Sf, SINEB2 flipped; ASf, Alu+SINEB2 flipped.

(c) A 73-bp overlap (OL) of antisense Uchl1 is sufficient to increase UCHL1 in transfected HEK cells. Scheme of mutant and scramble control in 5′ to 3′ orientation. Units for numbers along the left of gels in a–c indicate kDa. *(Figures are taken from Carrieri et al., 2012).*

Another important characteristic of the minimal 5' binding domain of *AS-Uchl1* is that the overlap region span across the translation initiation site (TIS or ATG) of sense mRNA, where the initial ATG start codon is centered with a −40/+32 configuration. Therefore, the *AS-Uchl1* overlaps to a portion of *Uchl1* mRNA 5′ UTR (untranslated region) and a portion of CDS (coding sequence) *(Figure 4.2a).* Interestingly, the artificial construct of a *(also referred as synthetic SINEUPS)* non coding sequence containing inverted SINEB2 element near to its 3' terminal and 5' sequence antisense to the ATG containing region of a target gene *(Figure 4.2b)* was also identified to exert translation upregulatory activity similar to that of *AS-Uchl1* (Zucchelli et al., 2015a). Although the role of ATG overlap, and the secondary structure of the target mRNAs around ATG remains unclear (Zucchelli et al., 2015a) in the functional activity of *AS-Uchl1* and SINEUPs, it has been suggested that the ATG overlap could contribute to provide high specificity of 5' binding domain of SINEUP, as it binds complementarily to the target mRNA. Additionally, the overlap within the 5' translated region (CDS) is identified as an important characteristic essential for SINEP like activity, as in the absence overlap to the CDS portion simply resulted to loss of SINEUP activity (Yao et al., 2015).

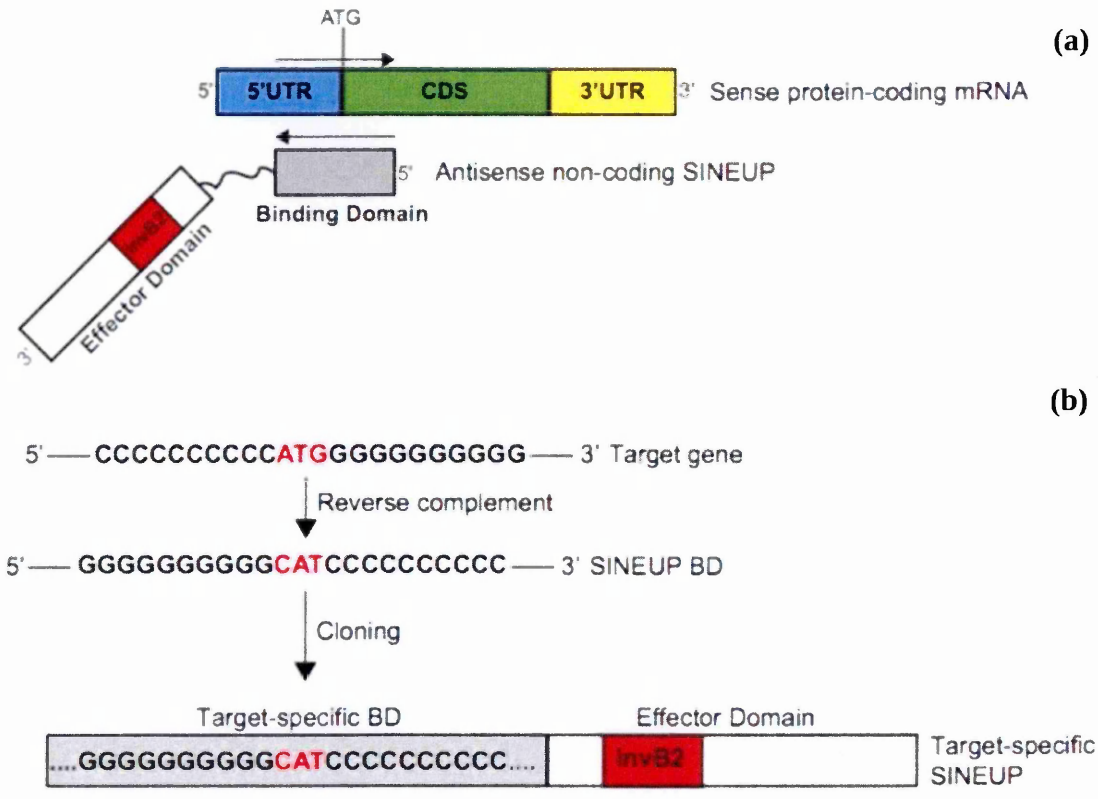


Figure 4.2 | Schematic representation of SINEUPs. (a) SINEUP modular structure. SINEUP binding domain (grey): SINEUP sequence that overlaps, in antisense orientation, to the sense protein-coding mRNA. SINEUP effector domain (red): non-overlapping portion of SINEUP (white), containing the inverted SINEB2 element (invB2) that confers activation of protein synthesis. Structural elements of protein-coding mRNA are shown: 5′ untranslated region (5′UTR, blue), coding sequence (CDS, green) and 3′ untranslated region (3′UTR, yellow). (b) Synthetic SINEUP design strategy. Schematic representation of the cloning strategy to generate target-specific SINEUPs. An artificial target gene sequence is indicated as example. *(Figure taken from Zucchelli et al., 2015a).*

The transcriptome-wide identification of S/AS transcript pairs using my pipeline revealed that only a small subset of sense coding genes have their initial ATG within ASlncRNA overlap region *(Figure 4.2).* For the sake of simplicity, from here on I would address the sense coding genes that have their initial ATG overlapped by ASlncRNAs as *"smRNA ATG"* genes and the sense coding genes that do not have their ATG overlapped by an ASlncRNAs as *"smRNA noATG"* genes.

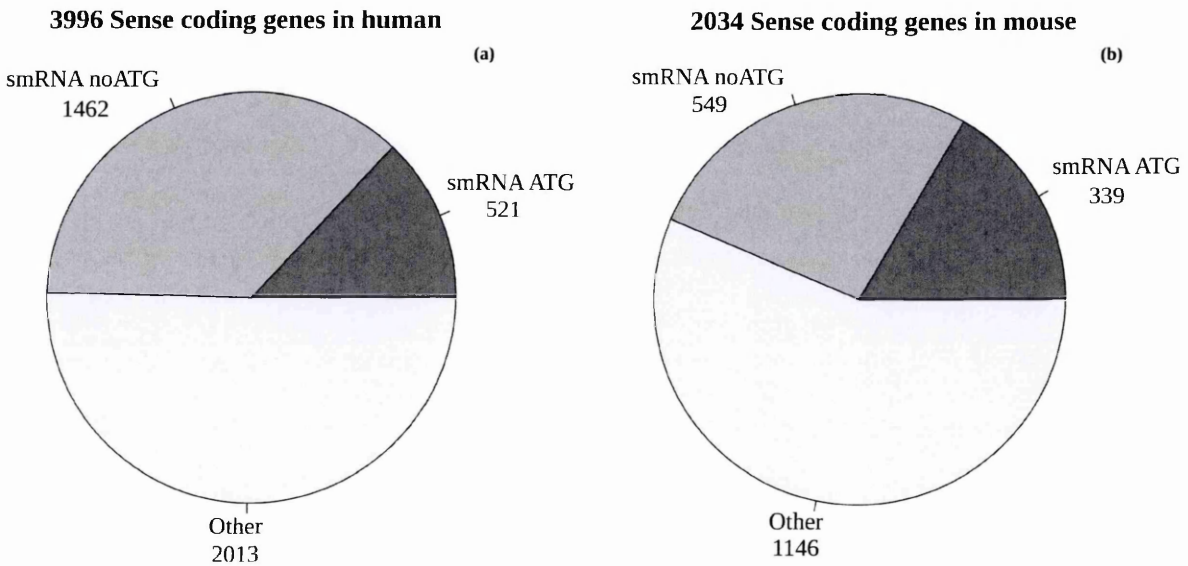


Figure 4.3 | Number of sense coding genes with or without ATG overlap. The pie represent the total number of identified S/AS gene pair in the transcriptomes of (a) human and (b) mouse. *smRNA ATG*, represents sense genes with at least one of its transcripts having an head-to-head overlap with an ASlncRNA where its ATG lie within the overlap region; *smRNA noATG*, corresponds to the head-to-head overlapping sense mRNA with none of its transcripts having the ATG within the

ASlncRNA overlap region; *other,* corresponds to other overlapping coding genes that are not in head-to-head overlap with ASlncRNAs.

In this chapter, I have described about several functional enrichment analysis I performed in order to determine the importance of ATG overlap among *smRNA ATG* genes and to understand if the *smRNA noATG* genes could behave similar to that of *smRNA ATG* gene set. In addition, I have also discussed about my investigation on how the ATG overlap characteristic of *smRNA ATG* genes could affect their translation. This is important to know because, if the canonical ATG start codon of an mRNA is overlapped by as ASlncRNA during cellular stress, it can't possibly take part into the translation initiation process thereby directly affecting the overall translation of *smRNA ATG* genes.

4.1.1. TIS switch hypothesis

if ASlncRNAs are functionally identical to *AS-Uchl1* then, in a given cellular stress condition they could get shuttled out from nucleus to cytoplasm, where they could bind to their respective sense coding mRNAs in a target-specific manner *(as described in case of AS-Uchl1)*. When an ASlncRNA binds to its target *smRNA ATG* gene, then the canonical ATG (TIS) of smRNA ATG gene gets blocked and becomes unavailable for translation initiation. In such an event, the translation initiation for *smRNA ATG* genes could occur from an alternative ATG present downstream (dATG) to the canonical ATG *(outside the ASlncRNA overlap region)*, where the internal ribosome entry site (IRES) would allow the ribosomes to initiate the process of translation (López-Lastra, Rivas, & Barría, 2005). The translation initiation from a dATGs could result into the production of variant proteins *(truncated protein, smaller in length)* that contain different NH2-terminal sequences in contrast to the proteins translated

from a canonical ATG. The NH2-terminal sequences of the proteins are usually known to carry signal peptides that control their sub-cellular localization (Choo, Tan, & Ranganathan, 2009), hence, the change of NH2-terminal sequence due to translation from a dATG could lead to the change in subcellular localization of the truncated protein. This implies that the *smRNA ATG* genes represent the gene set that corresponds to thedual localizing functional proteins which can switch their subcellular localization when their overlapping ASlncRNA is expressed.

To test this hypothesis, I performed functional enrichment analysis considering the annotation of function and the localization for each gene. Further, I performed an N-terminus signal peptide prediction analysis to determine the sub-cellular localization of the full-length and the truncated protein sequences corresponding to *smRNA ATG* in comparison to rest of the coding genes.

4.2. Materials and Methods

4.2.1. Functional enrichment analysis

For the determination of the biological functional annotations that are enriched or over-represented among *smRNA ATG* and *smRNA noATG* set of genes *(Figure 4.2),* in contrast to all protein coding genes, I performed a simple statistical proportion test using the *functional enrichment analysis module* of the pipeline. In order not to loose any bit of available GO related information in the enrichment analysis, I used the GO annotations corresponding to all evidence codes that included the manually curated and electronically assigned annotations. I performed this analysis taking into consideration of all 3 divisions of the GO annotations, i.e. the biological process (BP), molecular function (MF) and cellular component (CC), but successively focused on CC because the results corresponding to CC were most significant and similar between human and mouse. I implemented the *prop.test()* function in R to perform the 2-sample test (two-proportions test) for equality of proportions, where the first sample is one of the test gene groups under test *(smRNA ATG, smRNA noATG; coding genes with no antisense overlap Figure 4.2)* and the second sample is background gene list, i.e. the set of all annotated protein-coding genes. The *prop.test()* calculates a chi-squared statistic to test the null hypothesis, according to which the proportions of genes annotated to specific GO terms are same in the test and the background gene groups. The alternatives hypothesis in context to the analysis is that the proportion of genes annotated to a specific GO term is greater in the test gene group in contrast to background set of genes (Link to read more about *prop.test()* function https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.htm). Since multiple comparisons are performed in this analysis there is also a need for the correction of the obtained p-values to narrow down the chances of false discoveries. For this, once again the FDR method based p-value adjustment is performed using the *p.adjust()* function in R. In

order not be very conservative, only the test gene sets with a minimum of 15 annotated genes were considered for the p-adjustment step. The GO terms that are identified to be significantly (adjusted p-value < 0.08) over-represented into the test gene groups based on the above analysis, that are also common between human and mouse are selected for the representation into a comparative histograms representing the percentage of annotated genes for different test gene groups used in the analysis for further interpretations. Subsequently, to analyze a possible double localization of the protein products for different gene set, I prepared the combination of dual location GO annotations using the available multiple cellular component annotation for a single gene. And used the same to determine the over-representation of dual location annotations among all the test gene groups against the background list of genes, as explained above.

4.2.2. Prediction of N-terminus signals signal peptides

A signal peptide (sometimes referred to as signal sequence, targeting signal, localization signal, localization sequence, transit peptide, leader sequence or leader peptide) is a short (5-30 amino acids long) peptide present at the N-terminus of the majority of newly synthesized proteins that are either destined towards/inside certain organelles (the endoplasmic reticulum, golgi or endosomes) or secreted from the cell, or inserted into most cellular membranes. However, if the 5' CDS region of the mRNA is disrupted by an overlapping ASlncRNAs (among smRNA ATG genes) the fate of the so formed truncated protein could change because of the resultant changes in the N-terminus aminos acid sequence due to the overlap.

To get a clear picture of this, I decided to study and compare the N-terminus signal peptides among *smRNA ATG* against rest of the *smRNA noATG* genes and other coding genes with no

antisense overlap, because the proteins corresponding to *smRNA ATG* genes are more likely to have a disrupted N-terminus signal peptides due to the CDS overlap by ASlncRNAs, whereas this wont be the case in rest of the other coding genes, which make them a good control group. However, to proceed any further with this analysis it is important to deal with the redundancy of protein sequences, as each protein coding gene can have multiple alternatively spliced isoforms. For this, I considered to select only the longest coding isoform as the representative isoform for the prediction of N-terminus signal peptide among *smRNA ATG* genes and rest of the other genes. Additionally, I decided to specifically look for the N-terminus mitochondrial targeting signal peptides, because the functional enrichment analysis revealed *smRNA ATG* genes are significantly enriched for mitochondrial localization signals. For the identification of the N-terminus signals, I chose to use a previously published tool called ***targetp*** (Emanuelsson et al., 2007), which is able to predict if the N-terminus protein sequence contain a mitochondrial targeting signals peptide *(mTP)* or if the protein takes part in secretory pathways *(SP)* or gets localized into other cellular locations *(other)* considering the first 130 amino acids present in the N-terminus. The *targetp* is a neural network-based protein subcellular location prediction tool. For each of its prediction a neural network output scores are generated, which are not probabilities and do not necessarily add to one. However, by default the highest output score determines the prediction, hence the output scores are an indication of how certain a prediction is. Using these output scores, targetp generates reliability class (RC) score that ranges from 1 to 5. The RC is measure of the difference between the highest and the second highest output scores. For example, if this difference in the output scores is larger than 0.8, then the RC is determined as 1. Similarly if this difference is between 0.6 to 0.8 then RC is 2, and so on. The smaller RC score represents the more reliable predictions by *targetp* (Emanuelsson et al., 2000; Emanuelsson et al., 2007).

For the *targetp* prediction analysis, I used *-N* and *-c* parameters, where *-N* is to specify that the sequences being used in the prediction analysis are non-plant sequences and *-c* to specify to include the cleavage site of the signal peptides into the prediction. The protein sequences corresponding to *smRNA ATG*, *smRNA noATG* and rest of the genes with no ASlncRNA overlap are used as the input sequences for *targetp* prediction that are written into fasta format files. Further to test the enrichment or over-representation of the predicted signals (with special interest on mTPs) among *smRNA ATG* and *smRNA noATG* gene groups, in contrast to all protein coding genes with no ASlncRNA overlap, I performed a two sample proportion test using the *prop.test()* function in R. The null hypothesis tested here is that the proportion of genes with predicted signal peptides are same for the test gene groups (*smRNA ATG and smRNA noATGs*) and the background genes (coding genes with no antisense), whereas the alternative hypothesis is, the proportion of test genes containing signal peptides is greater than the background genes.

4.2.3. Identification of the change in N-terminus signal peptides between the full-length and truncated protein sequences

To test the hypothesis stated in *section 4.1.1*, I generated a list of full length and truncated protein sequences corresponding to *smRNA ATG* genes, where the full length sequences are the complete protein sequences and truncated sequences are the sequences starting from first dATG outside the overlapping region (*a minimum length of at least 10 amino acid sequence is considered*). The two set of sequences thus produced are analyzed for the presence of the N-terminus signal peptide using the *targetp*. The change in the predicted N-terminus signal peptide between the full length and truncated form for each protein sequence are recorded.

The total number of cases with a change in the N-terminus signal peptide before and after truncation for *smRNA ATG* genes, of course wont be very informative until it is compared against a control set. Hence, I generated a 100 random sample of protein sequences corresponding to rest of the genes (noASlncRNA and *smRNA noATG* genes) each with a sample size *n* = total number *smRNA ATG* genes. For each of thus generated random sample of protein sequences, I created a set of truncated sequences considering every time the observed number of nucleotides overlapped by an ASlncRNA in *smRNA ATG* gene set. Finally, the mean of total number of cases with the change in signal peptide between the full length and truncated random samples are comparable against the recorded number signal changes in case of *smRNA ATG* gene set. The identified percentage of genes with loss of a signal after the truncation for both *smRNA ATG* genes and random sample are then represented into charts for comparative interpretation.

4.2.4. Functional enrichment analysis considering SINE repeats and ATG overlap characteristics of ASlncRNA

In order to gain a combined insight of the cellular component (CC) associations of sense coding gene considering its overlapping ASlncRNA pair, it is important to spell out different *features* and *characteristics* of ASlncRNA and S/AS pair respectively that must be taken in to consideration. Here, *feature* of ASlncRNA refers to the two domains of a functional SINEUP–

1. 5' binding domain containing ATG overlap and
2. 3' effector domain with embedded SINE repetitive elements among their respective overlapping ASlncRNAs partners.

These two features of ASlncRNA can account for the following three characteristics of a S/AS pair of coding and ASlncRNA transcripts-

1. Overlap type *(ex: head-to-head or tail-to-tail as shown in Figure 2.2)*
2. ATG in overlap *(presence or absence of ATG overlap)*
3. SINE repeat and their specific orientation with respect to the ASlncRNA.

Based on the above mentioned features and characteristics of ASlncRNA and S/AS pair, all coding genes could be classified into multiple categories *(45 catagories; Table 4.1)* because a single gene can have multiple transcripts, each showing different characteristics of S/AS pair for a given feature of ASlncRNA. For example, a coding gene could have two or more transcripts with head-to-head overlap against an ASlncRNA, where one could have the ASlncRNA overlap spanning across its ATG, while other do not. At the same time the ASlncRNA transcript could contain multiple SINE elements that are either in inverted or direct orientation with respect to itself. As a second example, there could be a coding gene which has all of its transcript isoforms having their ATGs overlapped by an ASlncRNA transcript containing exclusively inverted SINE elements. To understand which of these characteristics of S/AS transcript pairs could be contributing towards a specific functional association of the sense coding gene, it becomes important to classify such genes into separate categories, lets say, *head.in-ATG-noATG.at-least-one-inverted-sine* and *head.ex-atg.ex-inverted-sine* for the above mentioned two examples respectively *(Figure 4.4)*.

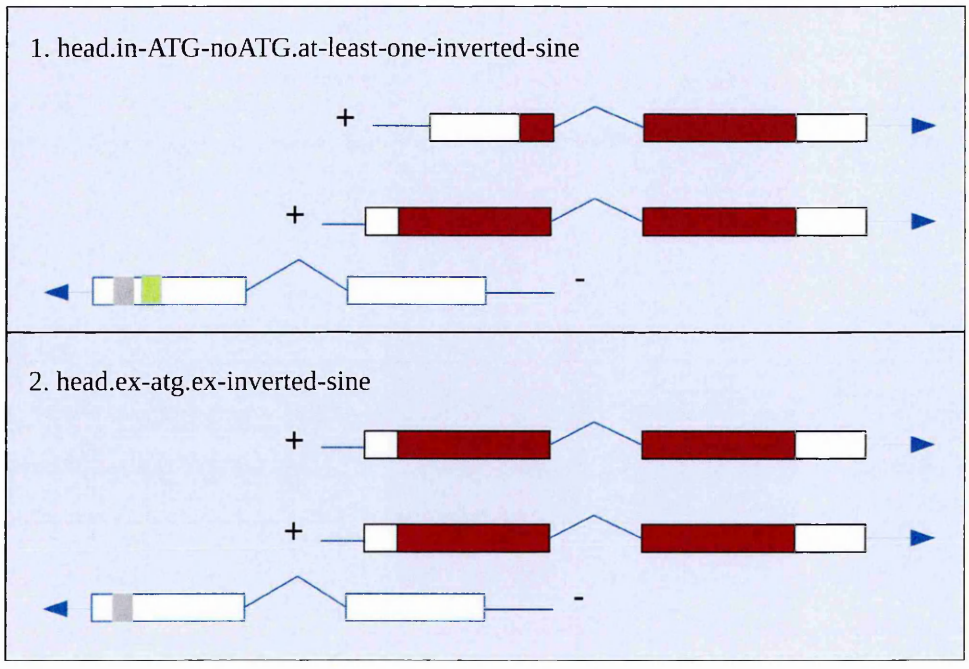


Figure 4.4 | Examples of sense coding gene categories. In the above charts boxes represents the exons which are connected by lines showing introns together they represent a transcript body. The strands of the transcripts are shown by plus (+) and minus (-) signs. Here, the transcripts on (+) strand are coding sense mRNAs where the shaded region of the boxes shows the CDS region and the unshaded region shows the 5' and 3' UTRs. Similarly, the unshaded boxes on (–) strands represents the exons for lncRNA transcripts, which contains gray and green boxes representing the inverted and direct oriented SINE repeats respectively.

Here, the name of each category is decided schematically. The "." symbol separates the three characteristics of a S/AS transcript pair (*overlap type, ATG overlap status, and SINE orientation),* whereas the "-" symbol is used to elaborate inside each characteristic to make it more readable. For example, considering the category name *head.in-ATG-noATG.at-least-one-*

inverted-sine, here, the "-" symbol is used to elaborate the second and third characteristics *(ATG overlap status and SINE orientations)* to explain that the ASlncRNA which is in *head-to-head* overlap with the sense coding gene is inclusive *(shown as "in" the name)* of both ATG overlap and no ATG overlap instances with respect to different transcript isoforms. Finally, the ASlncRNA contains at least one inverted SINE repeat. Similarly, considering the second category name mentioned above *head.ex-atg.ex-inverted-sine,* the "-" symbol is used to explain that the ASlncRNA which is in head-to-head overlap is exclusive *(indicated as "ex")* of ATG overlap instance with respect to all transcript isoforms of the coding gene and the ASlncRNA exclusively *(again shown as "ex" the name)* contains the SINE repeats in the inverted orientation. In cases where all kinds of overlap are considered such as *head-to-head, tail-to-tail, plus-inside* and *minus-inside (Figure 2.2)* the category name contains "all" in place of "head". The schema and the names of all 45 sub-categories of sense coding genes are shown in *table 4.1.* Here, it is also noticeable that the gene sub-categories with a characteristic of *"at-least-one-inverted-SINE"* are taken into consideration but not *"at-least-one-direct-SINE"* because the inverted SINE is identified as the effector domain in *AS-Uchl1.*

S. No	Sense coding gene categories	S. No	Sense coding gene categories
1	head.in-inverted-direct-sine	26	all.in-atg-no-atg.ex-inverted-sine
2	head.in-atg.no-atg	27	all.in-atg-no-atg.ex-direct-sine
3	head.in-atg-no-atg.sine	28	all.in-atg-no-atg.at-least-one-inverted-sine
4	head.in-atg-no-atg.ex-inverted-sine	29	all.ex-no-atg.without-sine
5	head.in-atg-no-atg.ex-direct-sine	30	all.ex-no-atg.sine
6	head.in-atg-no-atg.at-least-one-inverted-sine	31	all.ex-no-atg.ex-inverted-sine
7	head.ex-no-atg.sine	32	all.ex-no-atg.ex-direct-sine
8	head.ex-no-atg.ex-inverted-sine	33	all.ex-no-atg.at-least-one-inverted-sine
9	head.ex-no-atg.ex-direct-sine	34	all.ex-no-atg
10	head.ex-no-atg.at-least-one-inverted-sine	35	all.ex-inverted-sine
11	head.ex-no-atg	36	all.ex-direct-sine
12	head.ex-inverted-sine	37	all.ex-atg.without-sine
13	head.ex-direct-sine	38	all.ex-atg.sine
14	head.ex-atg.sine	39	all.ex-atg.ex-inverted-sine
15	head.ex-atg.ex-inverted-sine	40	all.ex-atg.ex-direct-sine
16	head.ex-atg.ex-direct-sine	41	all.ex-atg.at-least-one-inverted-sine
17	head.ex-atg.at-least-one-inverted-sine	42	all.ex-atg
18	head.ex-atg	43	all.atg.in-with-without-repeats
19	head.at-least-one-inverted-sine	44	all.atg.ex-without-repeats
20	all.no-atg.in-with-without-repeats	45	all.at-least-one-inverted-sine
21	all.no-atg.ex-without-repeats		
22	all.in-inverted-direct-sine		
23	all.in-atg.no-atg.without-sine		
24	all.in-atg.no-atg		
25	all.in-atg-no-atg.sine		

Table 4.1 | Sense coding gene categories. The category names contain "." symbol to separate the three characteristics, 1) overlap type, 2) ATG overlap and 3) SINE ostentations. While "-" symbol is used to internally elaborate each characteristic. *head* – head-to-head overlap; *all* – all possible S/AS overlap; *in* – inclusive; *ex* – *exclusive.*

Next, considering the above mentioned gene sub-categories as test gene sets, I computed the cellular component specific functional enrichment analysis with respect to all sense coding genes. The enrichment or over-representation of a CC among different test gene groups (Table 4.1) in contrast to all sense coding genes is determined using the two-sample proportion test, once again by implementing the *prop.test()* function in R. The null hypothesis which is tested in this case is that the proportion of test gene groups annotated for specific CC is same as that of the proportion of all annotated sense coding genes. Based on the p-values (<0.05) thus obtained, the test gene groups that show a significant over representation for the annotation of nucleus, cytoplasm and mitochondrion cellular components are are highlighted using a comparative chart representing the percentage of annotated genes and the p-values for further interpretation.

4.3. Results and discussions

4.3.1. Functional enrichment analysis for *smRNA ATG, smRNA noATG* and no antisense genes

The functional enrichment analysis considering all three sub-ontologies, cellular component (CC), molecular function (MF) and biological process (BP) for test gene lists categorized based on the presence or absence of ATG in the overlap region (*smRNA ATG* and *smRNA noATG*) and coding genes with no antisense overlap, with respect to background list of all protein coding genes (Proteome) revealed that *nucleus (GO:0005634), cytoplasm (GO:0005737)* and *mitochondrion (GO:0005739)* CC and *protein binding (GO:0005515)* and *DNA binding (GO:0003677)* MF GO terms are significantly enriched in different classes of test genes in human and mouse *(Figure 4.5)*. Here, it is important to keep in mind that in case of the CC ontology, the cellular components can share relationship among themselves, for example, considering nucleus, mitochondrion and cytoplasm, the first two are the part of the third. Hence the gene set enriched for cytoplasm might also contain a subset of genes that are annotated for nucleus or mitochondrion. Similarly, the nucleus and mitochondrion components do not share any part, hence the list of genes annotated to either of them could be mutually exclusive, unless a gene is annotated for both components (Link to learn more about ontology relationships http://geneontology.org/page/ontology-relations).

In the results, it is interesting to observe that the *smRNA ATG* class of genes are particularly enriched for mitochondrion in *both* human and mouse *(Figure 4.5 a,b)*. This suggests that the proteins encoded by *smRNA ATG* class of genes are more likely to be functional inside mitochondria, whereas the *smRNA noATG* class of genes are likely to encode proteins that are functional in nucleus or cytoplasm, although in case of human they are also enriched for

mitochondrion localization. The observed nucleus and cytoplasm specific enrichment of *smRNA noATG* genes in human and mouse is also supported by their enrichment observed for MF GO terms suggesting *smRNA noATG* encoded proteins are involved in DNA-binding in nucleus and protein-binding in cytoplasm *(Figure 4.5 c,d)*.

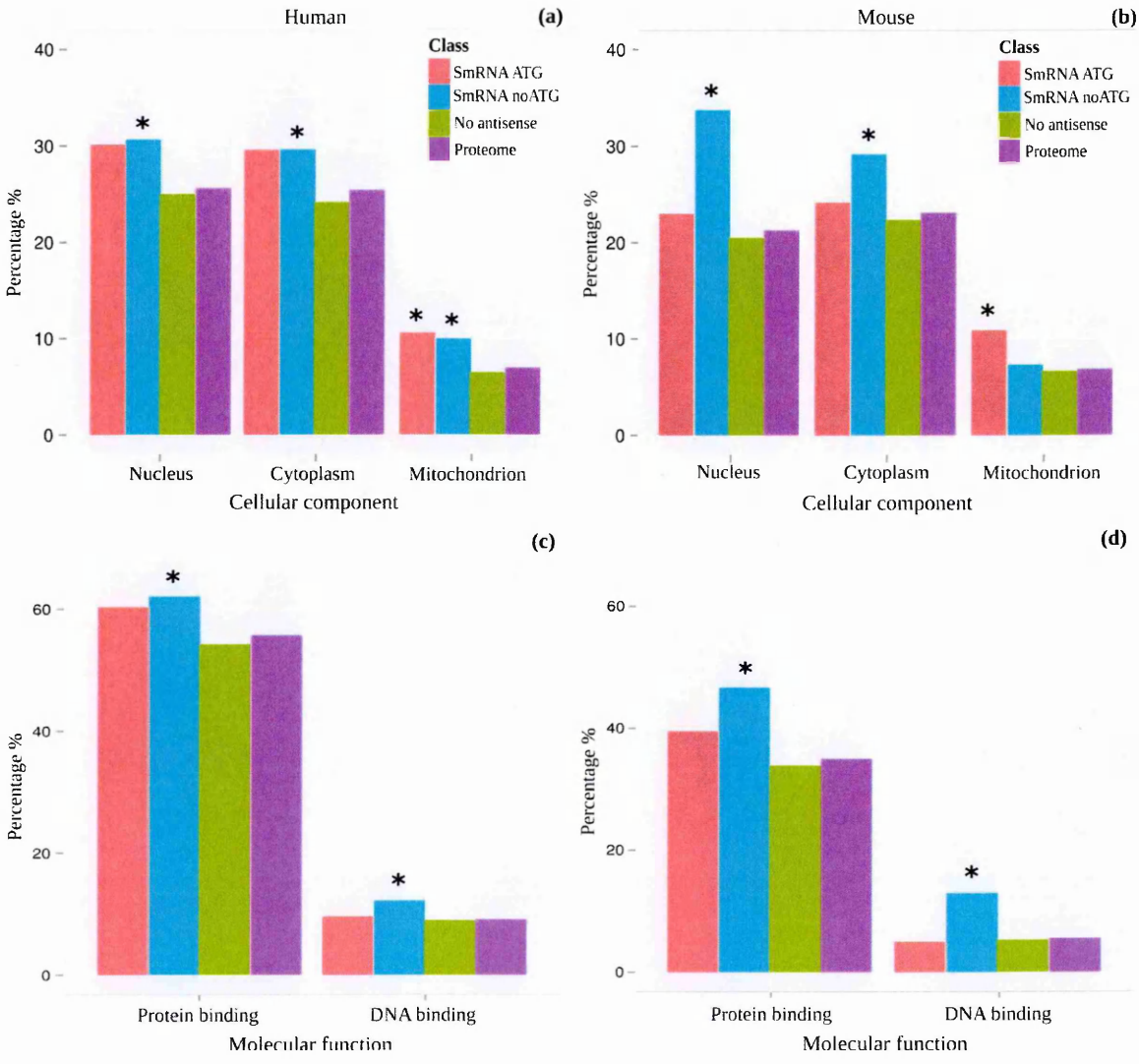


Figure 4.5 | Percentage of genes annotated for specific GO terms. In the above charts *y-axis* represents the percentage of gene groups annotated for specific GO terms belonging cellular components (a) and (b), and molecular functions (c) and (d) corresponding to human and mouse respectively. The significantly enriched gene classes (adjusted p-value < 0.08) for a given GO term are marked with (*) symbol.

Given that the *smRNA ATG* genes are specifically enriched for mitochondrial localization in human and mouse, one might expect that the "ATG overlap" characteristic among *smRNA ATG* genes could be responsible for mitochondrial localization of their resulting protein products and that these genes are mainly involved in the energy metabolism. However to know this, and to understand how the overlap of the initial ATG start codon by an ASlncRNA could affect the process of protein translation of sense mRNA, further investigations are needed.

4.3.2. Analysis of the dual localization functional enrichment for *smRNA ATG* genes

According to the hypothesis *(Section 4.1.1)*, during cellular stress the ASlncRNA binds to the target mRNA spanning across its TIS codon (ATG) resulting in the translation initiation from a dATG present outside the overlap region, thereby generating a variant protein with a change in its N-terminus sequence and the sub-cellular localization signal present on it. If this hypothesis is true, then *smRNA ATG* genes that are significantly enriched for mitochondrial localization would be translated from a dATG due to the overlap of ASlncRNA spanning across its TIS (expressed during cellular stress), thereby gaining the mitochondrial targeting signal in order to increase the specific functions of the mitochondrion. On the other hand In normal cell conditions, in the absence of ASlncRNA overlap, smRNA ATG genes might be functional in nucleus or cytoplasm. To test this hypothesis, I decided to perform a dual-localization functional enrichment analysis considering the following GO annotation combinations for mitochondrion and nucleus or other part of the cytoplasm -

1. *Nuc-Mit*, genes annotated for nucleus and mitochondrion
2. *Nuc-Mit/Cyt*, genes annotated for nucleus/mitochondrion/cytoplasm
3. *Nuc-Cyt*, genes annotated for nucleus/cytoplasm
4. *Mit-Cyt*, gene annotated for mitochondrion/cytoplasm

As already discussed in *4.3.1*, here again it is important to note that among the above mentioned four combination of cellular components, *Nuc-Mit*, is the only combination which represent mutually exclusive cellular components, hence the genes belonging to this category is necessarily annotated for both nucleus and mitochindrion representing the list of genes that are dual-localizing. However, the rest of the three categories contains the cellular components that could be the part of the one another and the genes belonging to these categories do not necessarily mean they are dually localizing. Hence, these categories are included in the functional enrichment analysis just for the sake of comparison between the *smRNA ATG, smRNA noATG and No antisense* genes, to find out if these gene set are enriched for Nuc-Mit dual-localization or rest of the other "possibly" dual localizing gene categories.

The functional enrichment analysis considering the dual-localization annotations as mentioned above for *smRNA ATG, smRNA noATG,* and *noASlncRNA* test genes against the proteome as the background list revealed that the *smRNA noATG* genes are only group of genes that are enriched for possible dual-location annotations such as NucMit/Cyt and Nuc-Cyt in human and mouse and Mit-Cyt in human, whereas *smRNA ATG* genes do not show any enrichment for dual-location annotations therefore negating my hypothesis *(Figure 4.6)*. Another important point which should be kept in mind while interpreting these results is that, even though the GO annotations types used in the analysis are corresponding to all evidence codes which includes the manually curated and electronically assigned annotations, only around 2 % of total genes in each category are identifed to have Nuc-Mit specific dually annotations. Moreover, the electronically assigned annotations are generally not considered as a highly reliable source of annotation in comparison to the manually curated evidence code as it

corresponds to the IEA (Inferred from Electronic Annotation) evidence code, that are not assigned by a curator. Hence, if the GO annotation corresponding to IEA evidence code were to be removed from my functional enrichment analysis that constitutes ~41% and ~44% of the total 84557 and 92031 available GO annotations for cellular components in human and mouse respectively, then the percentage of genes with Nuc-Mit dual annotations would be further less than 2% for all the analyzed gene categories. This indicates that although the GO annotations are very useful resource to have a high-level view of each of the three ontologies, it is not best suited to analyze at least the CC specific dual localization enrichment between different gene sets.

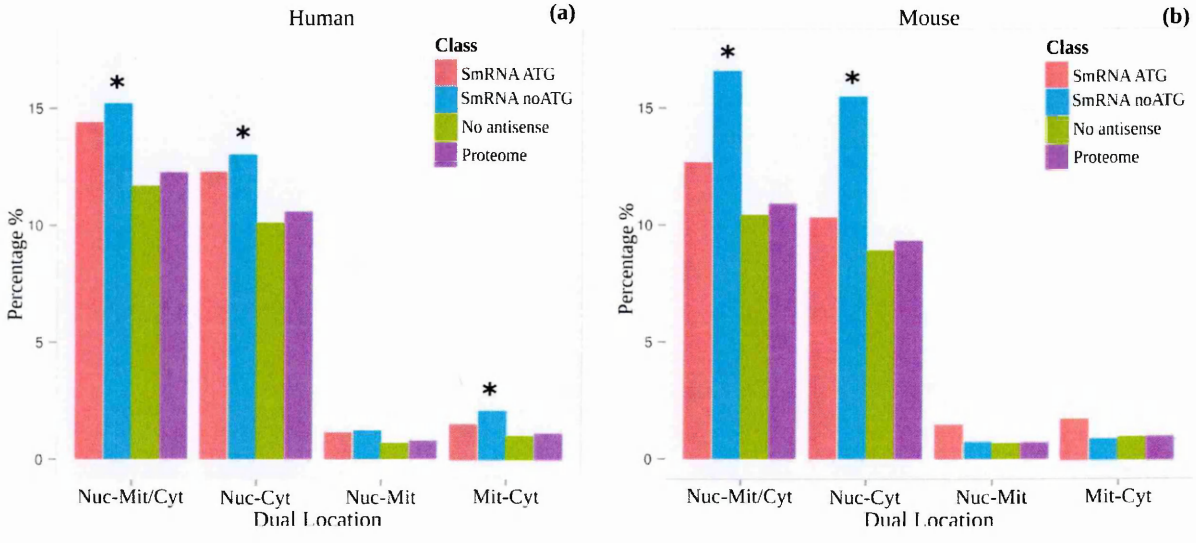


Figure 4.6 | Percentage of genes annotated for dual-locations. In the above charts *y-axis* represents the percentage of gene groups annotated for dual cellular locations in (a) human and (b) mouse. The significantly enriched gene classes (adjusted p-value < 0.08) for a given dual-location are marked with (*) symbol.

Based on the observation made by the dual-localization functional enrichment analysis, I next decided to test potential localization of the genes without relying on GO annotations but rather on the presence of specific signal peptide in the N-terminus of their protein products that are involved in the sub-cellular targeting of full-length and truncated proteins corresponding the mRNAs with and without ASlncRNA overlap spanning across their initial TIS receptively.

4.3.3. Analysis of the N-terminus signal peptides

The results of functional enrichment analysis previously showed that the *smRNA ATG* genes are significantly enriched for mitochondrial localization *(Figure 4.5)*. As a consequence, I further decided to look for the presence of mitochondrial targeting signal peptides (mTPs) in the N-terminus sequence of the full-length proteins, as the mTPs are known to be present in the N-terminal region of the protein sequence (Schatz & Dobberstein, 1996). For this purpose, I used a previously published tool called ***targetp*** (Emanuelsson et al., 2007) as discussed in methods ***4.2.2***.

The results of *targetp* predictions remained in agreement with the functional enrichment analysis. Because the *smRNA ATG* list of genes are identified to be significantly enriched for the presence of N-terminus mitochondrial targeting signal peptide with respect to noASlncRNAs, whereas, the smRNA noATG genes are identified to be enriched for the *other* localization signals *(Figure 4.7)*.

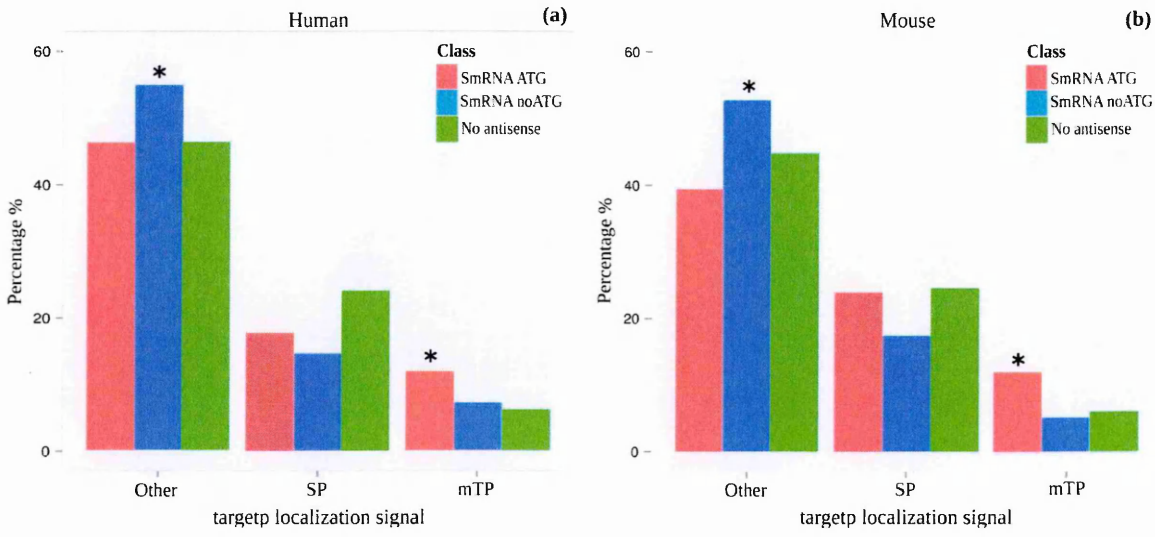


Figure 4.7 | Percentage of genes containing *targetp* localization signal. In the above charts y-axis represents the percentage of gene groups containing localization signals (RC <=3; reliability score generated by *targetp*) identified by *targetp* software in (a) human and (b) mouse. The significantly enriched gene classes (p-value < 0.08) for a given localization signal are marked with (*) symbol.

4.3.4. Analysis of the N-terminus signal transition from full-length to truncated protein sequences

Subsequently, I analyzed the transition of N-terminus signal peptides between full-length and truncated proteins, expecting a gain of mitochondrial targeting signals specifically among the truncated proteins. I first generated the list of truncated protein sequences for each full-length protein and performed the *targetp* signal peptide prediction analysis for both the full-length and truncated protein sequences and accounted for the signal transition, i.e, the change in sub-cellular targeting signals between the full-length and its truncated protein sequence. The

results revealed, majority of *smRNA ATG* genes contained a mitochondrial localization signal peptides in their full-length protein sequences which was lost in their truncated forms. In addition, many truncated proteins are identified to carry signal peptides that target the proteins to *other* locations but not mitochondria *(Table 4.2)*.

Species	Class	Predicted Signal	Before	After RC <= 3			Signal	
			RC <= 3	mTP	SP	Other	Present	Absent
Human	*smRNA ATG*	mTP	53	1	1	35	37	16
		SP	79	2	4	51	57	22
		Other	206	9	6	138	153	53
	100 random sample	mTP	28.98	1.07	2.03	18.85	21.95	7.03
		SP	108.3	3	15.67	61.06	79.73	28.57
		Other	215.17	4.44	9.61	152.18	166.23	48.94
			RC > 3					
	smRNA ATG	mTP	30	1	1	22	24	6
		SP	14	3	0	9	12	2
		Other	64	1	5	39	45	19
	100 random sample	mTP	29.48	0.97	2.03	19.32	22.32	7.16
		SP	20.19	0.42	6.2	8.71	15.33	4.86
		Other	58.48	1.67	5.56	37.18	44.41	14.07
Mouse			**RC <= 3**					
	smRNA ATG	mTP	34	0	0	26	26	8
		SP	68	0	9	39	48	20
		Other	112	3	6	79	88	24
	100 random sample	mTP	16.57	0.35	0.65	11.03	12.03	4.54
		SP	70.18	1.67	13.73	37.15	52.55	17.63
		Other	128.14	2.73	5.97	91.49	100.19	27.95
			RC > 3					
	smRNA ATG	mTP	18	0	0	8	8	10
		SP	14	0	1	8	9	5
		Other	38	0	4	25	29	9
	100 random sample	mTP	18.43	0.54	1.31	12.37	14.22	4.21
		SP	14.97	0.23	4.9	5.51	10.64	4.33
		Other	36.26	0.94	4.18	23.06	28.18	8.08

Table 4.2 | N-terminus signal transition from full-length to truncated protein sequence. The above table contains number of full-length *(column name "Before")*

and truncated proteins *(column name "After")* that are predicted to contain *mTP, SP or Other* signal corresponding to *smRNA ATG* genes and 100 random sample of genes (mean is shown in this case), for comparison *(column name "Class")*. Here, the number of proteins are filtered based on the Reliability Class (RC) scores which are generated by targetp for each prediction. Proteins with high reliable predictions (RC<= 3) are highlighted in green whereas the ones with lower reliable predictions are heightened in red (RC > 3). The second last column of the table represent total number of truncated proteins that are reliably (RC<=3) identified to posses a signal peptide *(column name "Signal present")*, whereas the last column contain total truncated proteins with less reliable predictions (RC>3) for containing any signal peptide. The truncated sequences are considered to loose signal peptides when the RC scores for the given prediction are greater than 3 *(column name "Signal absent")*.

Subsequently, to analyze if the observed signal transition is a specific characteristic of *smRNA ATG* genes I performed the randomization analysis as discussed in methods ***4.2.3.*** The randomization analysis revealed smRNA ATG genes do not behave different from random samples, in signal switch between the full length and truncated protein sequences *(Table 4.2).* As both random samples and smRNA ATG genes show a noticeably similar percentage of their truncated sequences to carry a signal peptide *(Figure 4.8).* The comparison of the number of cases with change in N-terminal signal peptides before and after the truncation between the smRNA ATG genes and 100 random samples disproves my hypothesis according to which the translation initiation from a dATG in *smRNA ATG* genes could result into a truncated protein that gains a different localization signal in its N-terminus sequence than the full length protein sequence. These results suggests that ATG overlap among *smRNA ATG* genes by an overlapping ASlncRNA is a characteristic that do not result to be involved in modulating the translation of mRNA in a way that it affect the localization of its protein products within the cell. Taken together, the role of ATG overlap in the S/AS pair of *Uchl1* mRNA, *AS-Uchl1* and synthetic SINEUPs remains elusive. Currently, the only available explanation for the importance of ATG overlap is suggested by *Yao et al., 2015.* according to whom the ATG overlap by an ASlncRNA could be a feature of ASlncRNA that is simply involved in increasing the the specificity of ASlncRNA for the targeted binding to mRNAs.

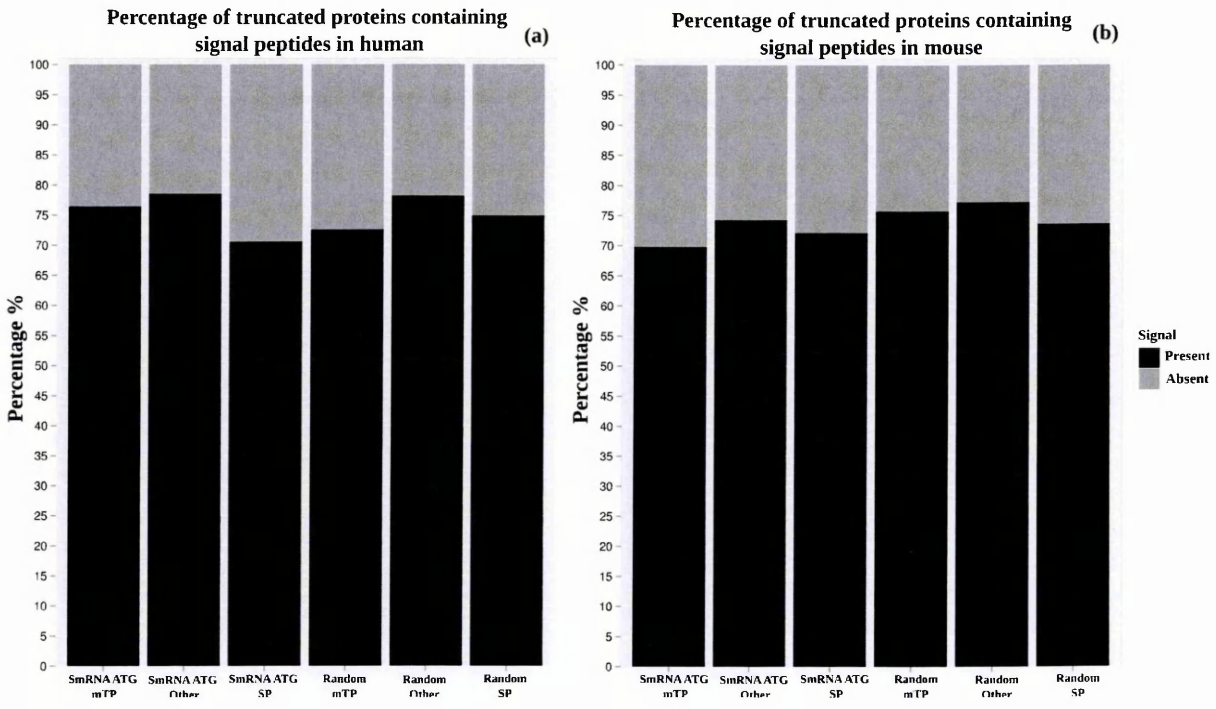


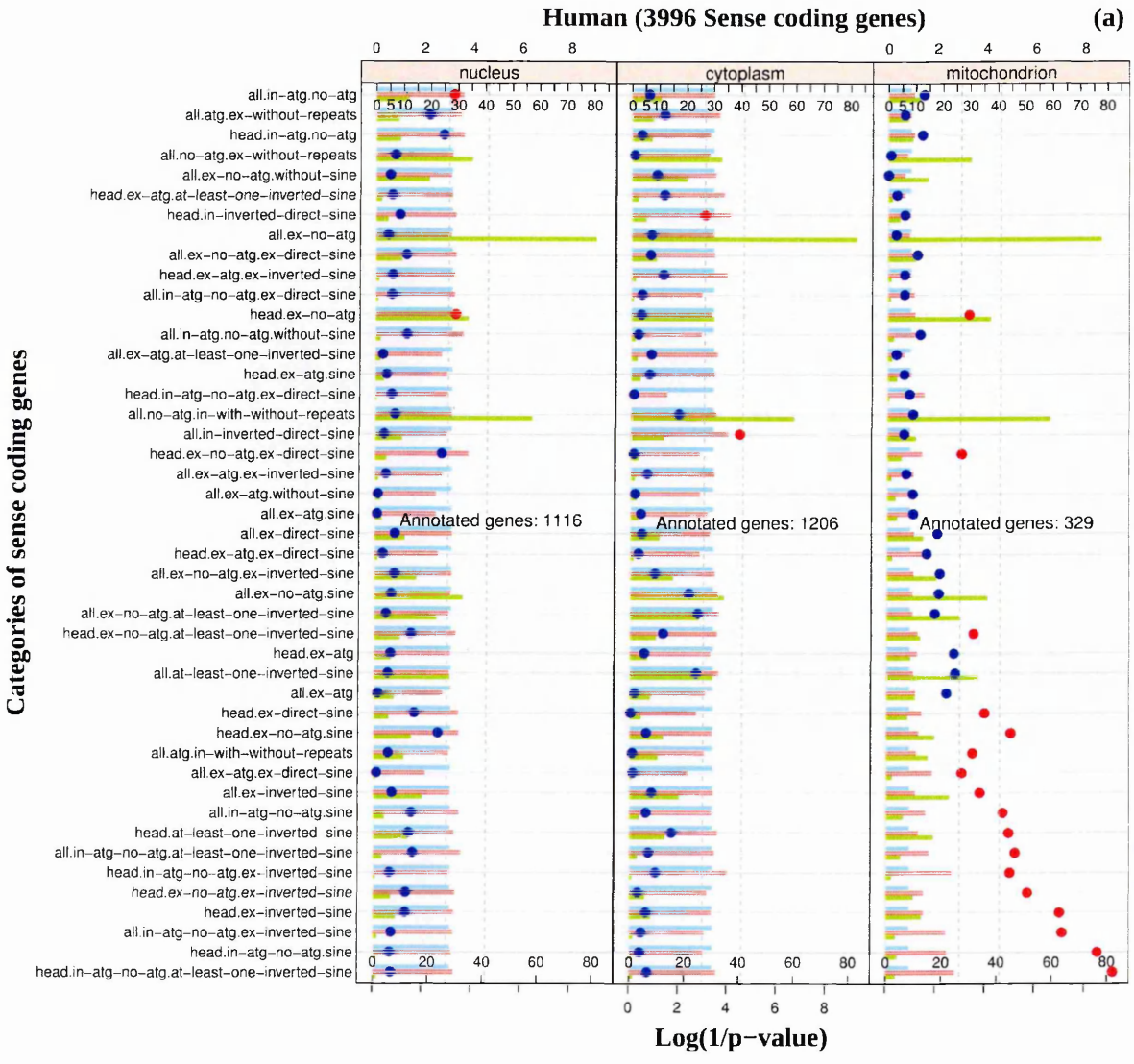
Figure 4.8 | Percentage of truncated proteins containing a signal peptide. The above charts *y-axis* represents the percentage of truncated proteins with or without a predicted signal peptides among *smRNA ATG* and 100 random samples for (a) human and (b) mouse.

4.3.5. Functional enrichment analysis for sense coding genes considering SINE repeats in combination with ATG overlap characteristics.

Until now, I have tried to analyze the importance of the ATG overlap characteristics of the 5' binding domain of ASlncRNA over the functional associations of the sense coding genes. I identified that the sense coding genes with ATG overlap are significantly enriched for mitochondrial localization. However, dual localization analysis revealed inconclusive results and the N-terminus signal peptide prediction analysis revealed the ATG overlap by ASlncRNA

is not likely to affect the sub-cellular localization by TIS switch. Hence, I decided to consider another important module of functional SINEUPs which is its 3' effector domain that contains an inverted SINEB2 repetitive element, and look for functional enrichment as described in section **4.2.4** to gain a more complete perspective of the CC specific functional associations *(focusing mainly on nucleus, cytoplasm and mitochondrion)* of sense coding genes in regard to ATG overlap along with SINE repeats and their specific orientation within ASlncRNAs. This analysis takes into account both features of the S/AS pairs, i.e. the ATG overlap and the presence and orientation of a SINE element.

Interestingly, the result of this analysis revealed, majority of sense coding gene categories that are significantly enriched for mitochondrion annotations belonged to *head-to-head* overlap type class, where the overlapping ASlncRNA contained either, at least one inverted SINE or exclusively inverted sine repeats in both human and mouse. However, human and mouse differed in terms of the ATG overlap characteristics. For example, mitochondrion specific enriched sense coding genes in mouse showed exclusive ATG overlap instances which means all transcript isoforms of these genes showed ATG overlap, whereas in case of human the sense coding genes showed the inclusive instances of ATG and no ATG overlap, where some if the transcript isoforms showed ATG overlap by ASlncRNAs and some did not *(Figure 4.9a and b).* This observation remained consistent with the results of previous functional enrichment analysis performed using *smRNA ATG* and *smRNA noATG* group of genes against complete proteome *(Figure 4.4).* Finally, as already observed, the nucleus specific enrichment of sense coding genes with no ATG overlap could be seen in case of mouse but not in case of human *(Figure 4.5 and Figure 4.9).*


Human (3996 Sense coding genes)
(a)
nucleus
cytoplasm
mitochondrion
Annotated genes: 1116
Annotated genes: 1206
Annotated genes: 329
Categories of sense coding genes
Log(1/p−value)
all.in−atg.no−atg
all.atg.ex−without−repeats
head.in−atg.no−atg
all.no−atg.ex−without−repeats
all.ex−no−atg.without−sine
head.ex−atg.at−least−one−inverted−sine
head.in−inverted−direct−sine
all.ex−no−atg
all.ex−no−atg.ex−direct−sine
head.ex−atg.ex−inverted−sine
all.in−atg−no−atg.ex−direct−sine
head.ex−no−atg
all.in−atg.no−atg.without−sine
all.ex−atg.at−least−one−inverted−sine
head.ex−atg.sine
head.in−atg−no−atg.ex−direct−sine
all.no−atg.in−with−without−repeats
all.in−inverted−direct−sine
head.ex−no−atg.ex−direct−sine
all.ex−atg.ex−inverted−sine
all.ex−atg.without−sine
all.ex−atg.sine
all.ex−direct−sine
head.ex−atg.ex−direct−sine
all.ex−no−atg.ex−inverted−sine
all.ex−no−atg.sine
all.ex−no−atg.at−least−one−inverted−sine
head.ex−no−atg.at−least−one−inverted−sine
head.ex−atg
all.at−least−one−inverted−sine
all.ex−atg
head.ex−direct−sine
head.ex−no−atg.sine
all.atg.in−with−without−repeats
all.ex−atg.ex−direct−sine
all.ex−inverted−sine
all.in−atg−no−atg.sine
head.at−least−one−inverted−sine
all.in−atg−no−atg.at−least−one−inverted−sine
head.in−atg−no−atg.ex−inverted−sine
head.ex−no−atg.ex−inverted−sine
head.ex−inverted−sine
all.in−atg−no−atg.ex−inverted−sine
head.in−atg−no−atg.sine
head.in−atg−no−atg.at−least−one−inverted−sine

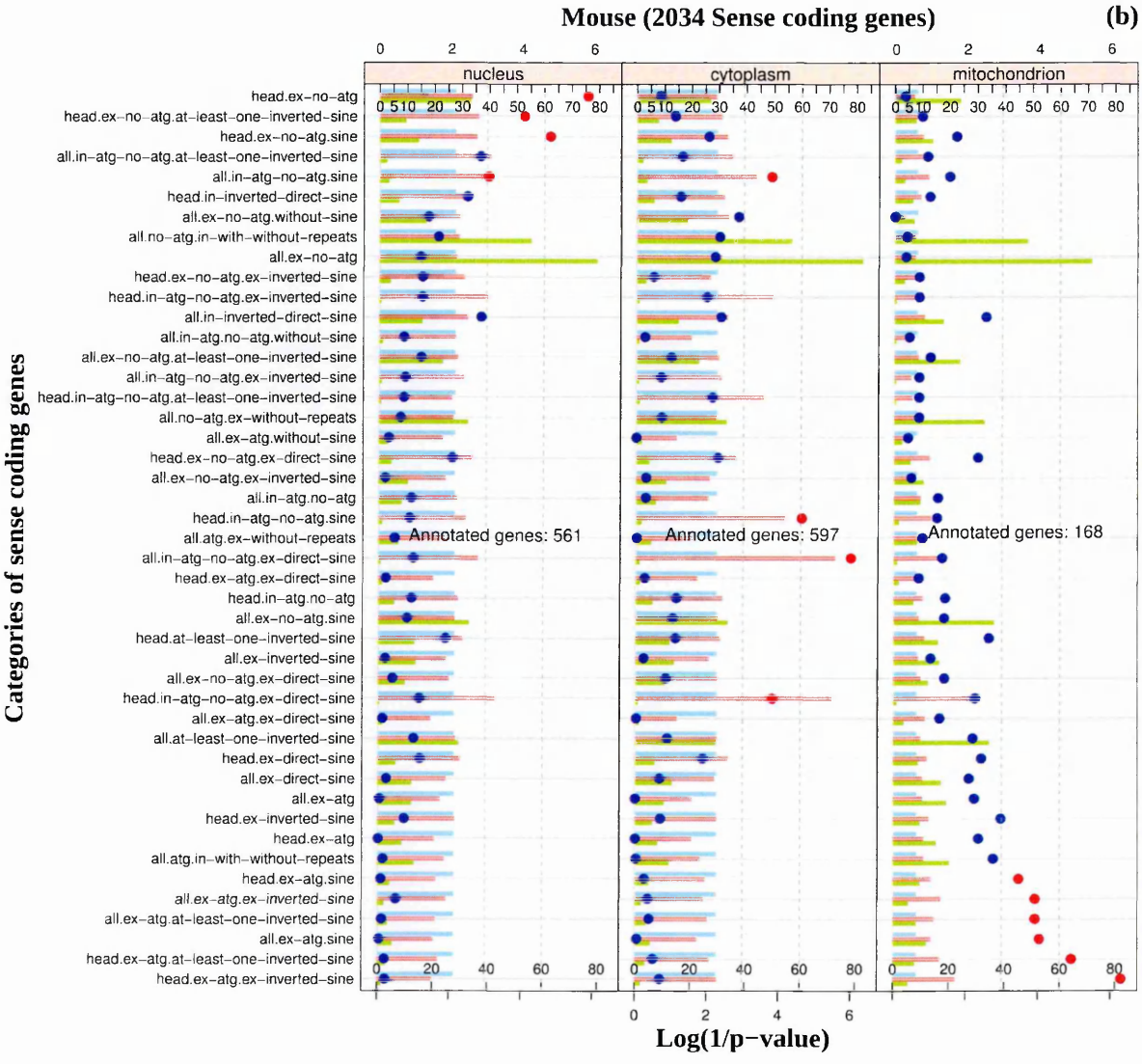


Figure 4.9 | CC specific functional enrichment for sense coding genes. The above charts represents the functional enrichment results of 45 sense coding gene categories listed in the left-hand side label-grid for nucleus/cytoplasm and mitochondrion shown in separate panels side by side. Two scales could be seen on the top and bottom of each panels. The one shown at the bottom of the panel represent the Log(1/p-value) and corresponds to the blue and red dots. Red dots are

used to show significant p-values. The vertical dotted lines are used to indicate p-value cut offs ranging from 0.05 (left) and 0.01 (right). The scale on top of each panel represent the percentage of annotated genes, the light blue colored bar represent the percentage of annotated genes in the background list (all sense coding genes) whereas, the red colored bar correspond to the percentage of annotated genes in the test gene list (gene categories in the left hand grid). Finally, the green colored bar represents the percentage of annotated test genes out of the total annotated background genes hence giving an indication about the power of test.

4.4. Conclusions

The results of the analysis described in this chapter indicate that sense coding genes with ATG overlap are significantly enriched for mitochondrial localization in human and mouse, whereas the sense coding genes with no ATG overlap are significantly enriched for nucleus/cytoplasm localization This observation is also supported by the enrichment analysis performed using the N-terminus signal peptides predicted using *targetp*. The signal peptide prediction analysis also revealed, the ATG overlap is unlikely to be involved in modulating the translation of mRNA by TIS switch, in a way that can affect the localization of the resultant protein products within the cell. Hence the importance of ATG overlap still remains unclear for a S/AS pair of transcripts, when the ASlncRNA is an effective SINEUP. Currently, the only existing explanation on the importance of ATG overlap in a S/AS pair is suggested by *Yao et al., 2015* according to whom the ATG overlap increases the specificity of ASlncRNA in targeted binding with mRNAs.

The functional enrichment analysis of multiple sub-categories of sense coding genes classified considering the presence of SINE repeat and the ATG overlap revealed that the majority of genes groups that were enriched for mitochondrial localization, had *head-to-head* overlap and commonly contained at least one inverted SINE element in their overlapping ASlncRNAs in both human and mouse. However, human and mouse behaved differently when we look for the ATG overlap characteristics of sense coding genes that are annotated for mitochondrion. Because in case of mouse the sense coding genes annotated for mitochondria have all their transcript isoforms overlapping to ASlncRNA spanning across the initial ATG codon. Unlike mouse in human, the mitochondria annotated sense coding genes contain a fraction of

transcript isoforms that have their ATGs overlapped by an ASlncRNA and a fraction that overlap to the ASlncRNA only at their 5' UTR region.

An important point which should be noted here is that all the analysis described in this chapter greatly relies on the precision of TIS annotations within the transcript models available from Ensembl/Havana, because based on this sole factor the gene groups used in my analysis are classified and compared. If we look at the transcripts structure of the *Uchl1* gene in the Ensembl genome browser *(Figure 4.10)*, we could observe in both the species that *Uchl1* show the presence of annotated ASlncRNAs containing an inverted SINE at their 3' end. Another important point which could be observed is that in mouse the S/AS overlap encompasses the region around TIS while this is not seen in case of human. We have seen and discussed the importance of ATG overlap in case of mouse *AS-Uchl1* function. But what about *AS-UCHL1* in human? Does the absence of the ATG overlap between human *UCHL1* and *AS-UCHL1* mean that human *AS-UCHL1* behaves different from mouse *AS-Uchl1? We* do not know it yet. Although based on various analysis described in this chapter, we witnessed that the sense coding genes with or without ATG overlap in human behave similar to the sense coding genes exclusively with ATG overlap in mouse. However, we can not be sure about the effects of the absence of ATG overlap in human *AS-UCHL1* function. Further wet-lab experiments aiming to understand the importance of ATG overlap in S/AS pair of genes would be required to shed more on this topic.

Interestingly, If we have look at specifically the brain derived gene models for *UCHL1* taken from *Human body map 2.0* (dataset not included in my study) for human *(Figure 4.10a)*, we could see that the annotated transcript show a different exon/intron structure, from that of the

Ensembl *UCHL1* gene model and the TIS is annotated within the ASlncRNAs overlapped region, however, similar comparison between the mouse *Uchl1* gene model from Ensembl and mouse brain RNASeq derived gene model (dataset not included in my study) remained consistent *(Figure 4.10b)*. This suggests that the TIS annotations may vary, especially in humans because the data from neural samples are more difficult to obtain. Therefore, I decided to ignore the ATG overlap based classification criteria in rest of my analysis and focus mainly on the SINE repeat content and orientation based gene classifications.

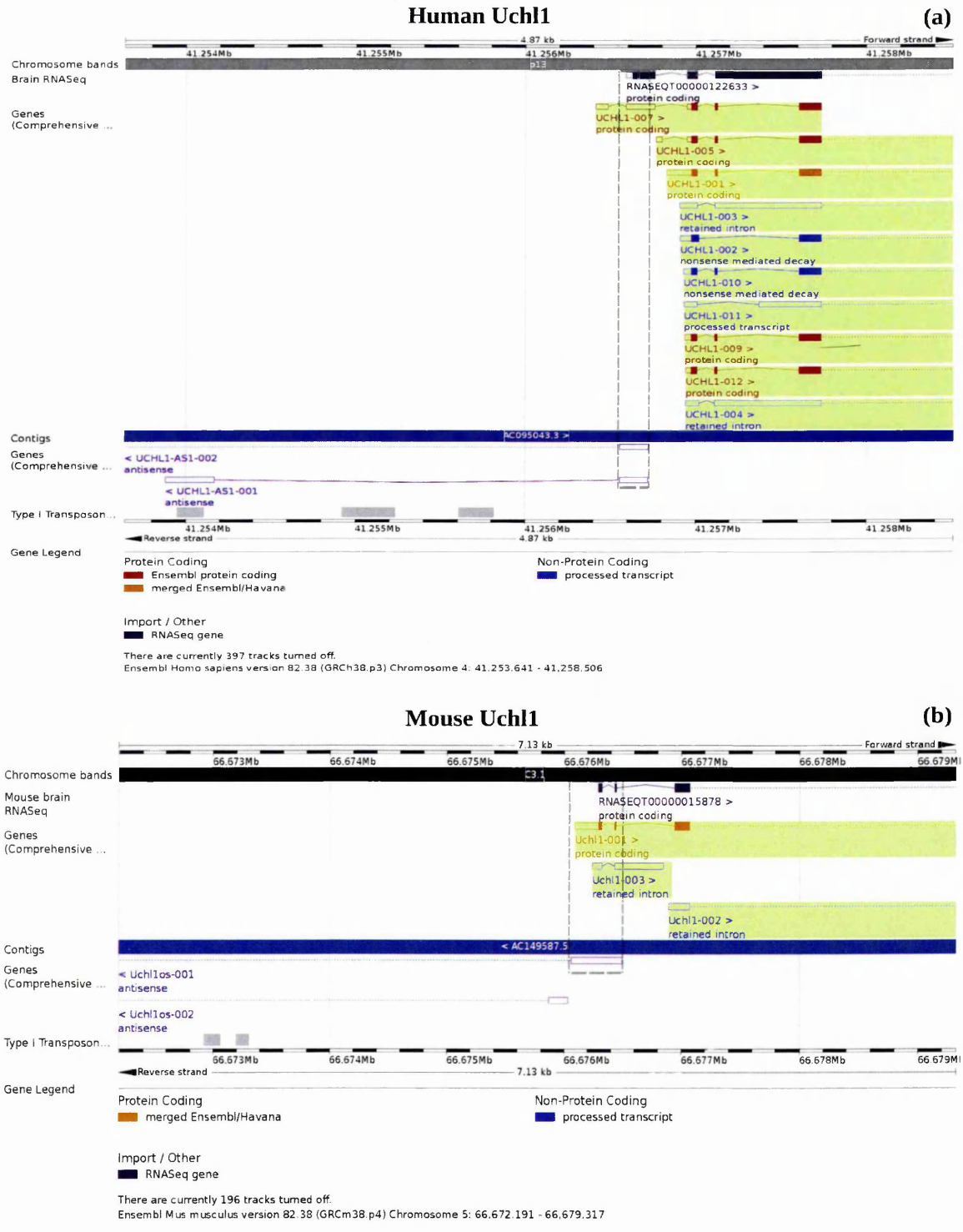


Figure 4.10 | Screen-shots of Ensembl browser showing the chromosome location containing *Uchl1* gene in (a) human and (b) mouse. The charts represent

Chapter 5

Analysis of the effect of SINE orientation on the functional activity of ASlncRNAs

5.1. Introduction

The effector domain in *AS-Uchl1* and synthetic SINEUPs is represented by an inverted SINEB2 element which is embedded near the 3' end of the transcripts and is necessarily required for their post-transcriptional protein up-regulatory activity *(Figure 1.2, 4.1)*. Although the natural *AS-Uchl1* in mouse contain a *direct Alu* along with the inverted SINEB2 repetitive element, the *Alu* was not found to be involved in the protein up-regulatory activity by *Carrieri et al. (Section 4.1)*. Interestingly, functional enrichment analysis of sense coding gene sub-groups classified based on the characteristics of ATG overlap and the orientation of SINE elements in their ASlncRNA revealed, the genes overlapping ASlncRNA which contain exclusively inverted SINE or at-least one inverted SINE are significantly enriched for mitochondrial localization in human and mouse. The results suggest a possible involvement of ASlncRNAs with embedded inverted SINE on the definition of the sub-cellular localization for their respective sense coding genes. However, in order to identify functional associations of sense coding genes purely based on the presence of specific SINE orientations in their ASlncRNAs, it becomes important to classify them into separate categories considering only SINE orientations in ASlncRNA as shown in *figure 5.1*.

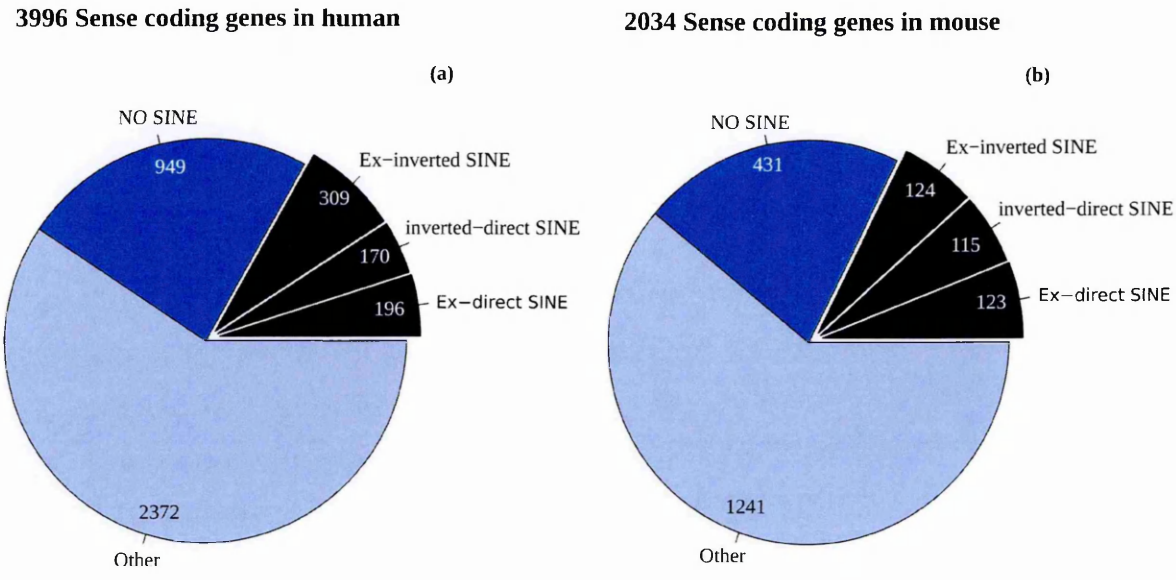


Figure 5.1 | Classification of sense coding genes. Pies represent the total number of coding genes with an antisense overlap in human (a) and mouse (b). Ex-inverted SINE, inverted-direct SINE, Ex-direct SINE represents the number of sense coding genes with ASlncRNAs *(head-to-head organization)* whose isoforms contain respectively exclusively inverted SINE, both inverted and direct SINE and exclusively direct SINE. NO SINE, represents sense genes with an ASlncRNAs containing no SINE repeats *(head-to-head or tail-to-tail or internal organization)*. Other, represents the remaining genes with ASlncRNA overlap *(tail-to-tail or internal organization)*.

In this chapter, I have described the functional enrichment analysis using the above mentioned gene categories in order to identify functional associations of sense coding genes based on the specific orientations of SINEs in their respective ASlncRNAs. Additionally, I have also discussed about a specific analysis I performed in order to understand, how sense coding gene

categories would behave during normal and cellular stress conditions. For this, I used the data from a published study, on RNAs levels associated to different polysome fractions in human MRC5 cell lysates in control and oxidative stress conditions (Giannakakis et al., 2015). Finally, I have described about the enrichment analysis for the 5'-TOP (terminal oligopyrimidine tract) motifs among the sense coding gene classes *(Figure 5.2)* in order to analyze their associations with the mTORC1 stress signaling pathway. The 5'-TOP are the motifs that are first identified to be found in mRNA transcripts for all ribosomal proteins studied to date, as well as in the protein synthesis elongation factors. These are present next to the 5' terminal cap structure and starts with a cytosine, which is succeeded by a stretch of 5–14 pyrimidines (Jefferies et al., 1997). In later studies TOP motifs were reported to be present in a wider variety of genes that code for lysosome and metabolism related proteins. They are also proposed to play an important role in the gene expression controls among the majority of cellular mRNAs (Yamashita et al., 2008).

As already discussed in section ***1.5.2.3***, *AS-Uchl1* is shown to be associated with mTORC1 *(mechanistic target of rapamycin, complex 1)* signaling pathway, where its inhibition by rapamycin *(a drug targeting mTORC1)* is shown to facilitate the shuttling of *AS-Uchl1* from nucleus to cytoplasm, where it can exert its function of up-regulating the translation of *Uchl1* mRNA thereby increasing UCHL1 protein levels (Carrieri et al., 2012). The mTORC1 is a protein complex that promotes critical cellular processes such as protein synthesis by controlling the phosphorylation of the regulators of translation such as p70S6K (p70 ribosomal S6 kinase) and translational repressor, eukaryotic initiation factor 4E-BP (eIF4E-binding protein). The phosphorylation of the first, facilitates the assembly of eIF3 (eukaryotic initiation factor 3) translation initiation complex, while the phosphorylation of second leads to its

disassociation from eukaryotic initiation factor 4E (eIF4E) *(Figure 5.2)*, allowing eIF4E to initiate cap-dependent translation (Ekim et al., 2011; Cargnello et al., 2015).

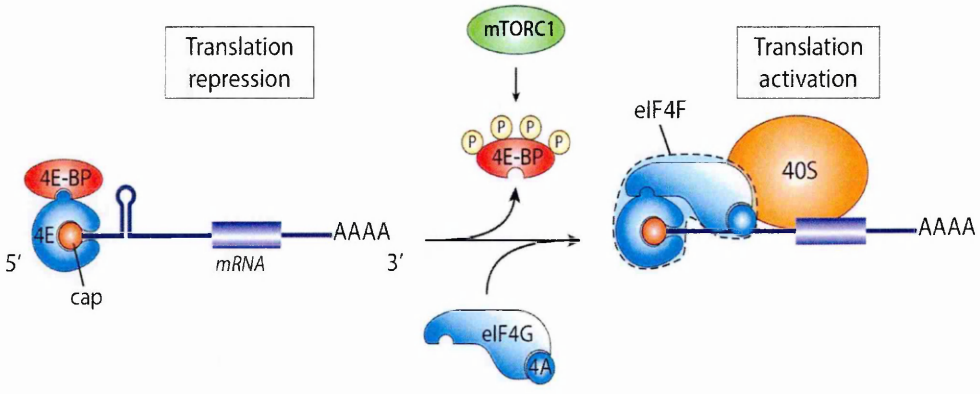


Figure 5.2 | Role of mTORC1 in protein synthesis. In quiescent cells, 4E-BP is hypophosphorylated and tightly associated with eIF4E, thus preventing translation initiation. When activated, mTORC1 phosphorylates 4E-BP leading to its dissociation from eIF4E and assembly of the eIF4F complex. 4E-BP repression by mTORC1 stimulates global protein synthesis. *(Above figure is taken from Cargnello et al., 2015).*

Interestingly, the work of *Thoreen et al.*, in 2012, showed that a subset of mRNAs that are specifically regulated by mTORC1, consists almost entirely of transcripts with established 5'-TOP motifs. Also, the inhibition of mTOR influences the mRNA translation that are mainly mediated by 4E-BPs, wherein a moderate suppression of the translation of all mRNA is seen, but a more marked inhibition were noted particularly in the case of TOP and TOP-like mRNA translation. Based on their study *Thoreen et al* also proposed a simple model that explains how mTORC1 differentially controls the translation of specific mRNAs (A schematic

representation is shown in *figure 5.4*). According to this model 4E-BPs inhibit translation initiation by interfering with the interaction between the cap-binding protein eIF4E and eIF4G. Loss of this interaction diminishes the capacity of eIF4E to bind TOP and TOP-like mRNAs much more than other mRNAs, there by selectively suppressing their translation (Thoreen et al., 2012).

However, *Uchl1* mRNA was shown to make an exception in this case. *Carrieri et al.*, showed that the inhibition of mTORC1 with the rapamycin treatment, although led to a slight impairment of the global translation with a noted depophosphorylation of 4EFBP and p70S6K *(Figure 5.3 a),* the *Uchl1* mRNA showed a marked increase in translation which was mainly regulated by the overlap of the modular *AS-Uchl1* that contains an inverted SINEB2 repeat.

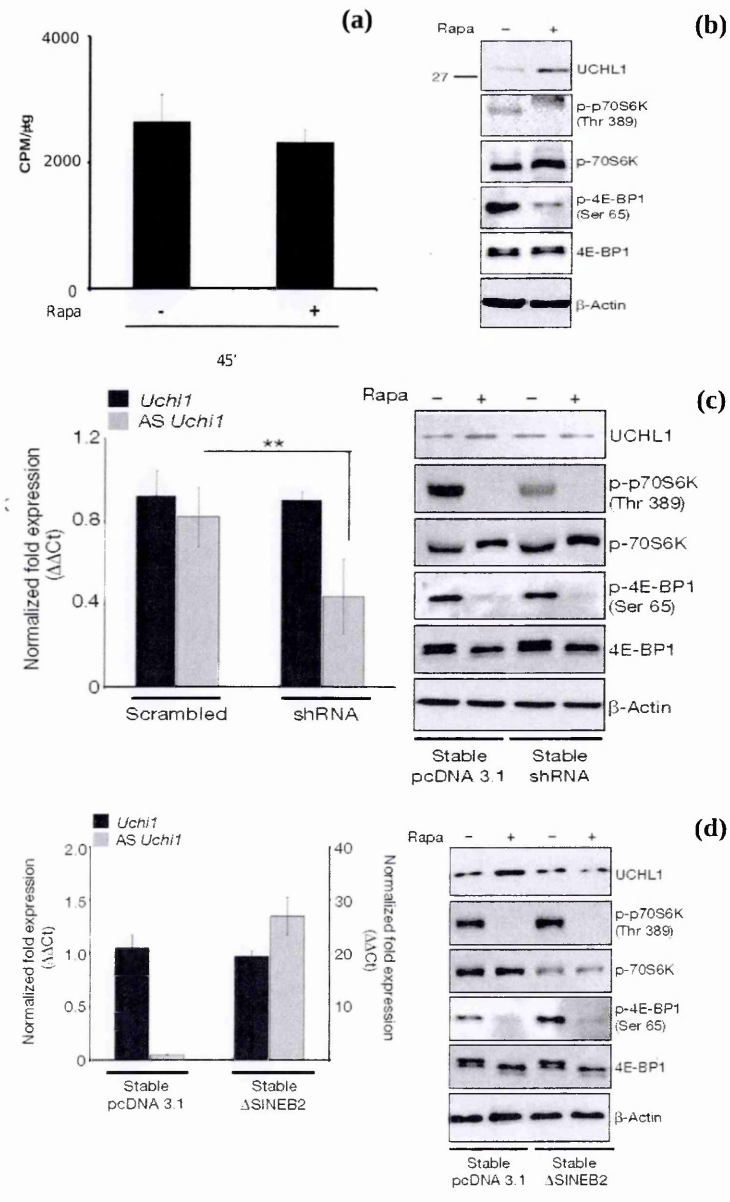


Figure 5.3 | *AS-Uchl1* mediates UCHL1 protein induction by rapamycin. (a) mTORC1 inhibition by Rapamycin (Rapa) treatment (+) for 45 minutes *(x-axis)* slightly impairs the global rate of translation *(y-axis)* in comparison to DMSO control (-). (b) UCHL1 protein level is increased in rapamycin-treated MN9D cells. Rapamycin inhibition of mTOR pathway is verified with anti-p-p70S6K and anti-p-

4E-BP1 antibodies. B-Actin is used as control. (c) Silencing *AS-Uchl1* transcription (shRNA) in MN9D cells inhibit rapamycin-induced UCHL1 protein level. Left, mRNA levels; right, protein levels. (d) Deletion of embedded SINEB2 (ΔSINEB2) is sufficient to inhibit rapamycin-induced UCHL1 protein upregulation *(Figures are taken from Carrieri et al., 2012).*

5.1.1. Deprived 5' TOP motif hypothesis

Considering that the 5'TOP motifs are found in diverse set of mRNAs and are not just limited to mRNAs coding for ribosomal proteins (Yamashita et al., 2008). Also at the same time, looking into the mTORC1-dependent translation control model proposed by *Thoreen et al., 2012* and the *Uchl1* mRNA translation regulation by *AS-Uchl1* explained by *Carrieri et al., 2012 (Figure 5.4),* I hypothesized that the sense coding genes overlapping to ASlncRNAs containing inverted SINEs that are likely to act as *AS-Uchl1,* could be deprived of 5'TOP motifs. This is because, their translation regulation are under the control of *AS-Uchl1* like ASlncRNAs and independent of the 5'TOP motifs. To test this hypothesis bioinformatically, I performed an enrichment analysis (described later in this chapter) to test the over-representation of TOP motifs in sense coding genes overlapping to ASlncRNAs inverted SINE repeats in contrast to sense coding genes overlapping to ASlncRNAs without any SINE.

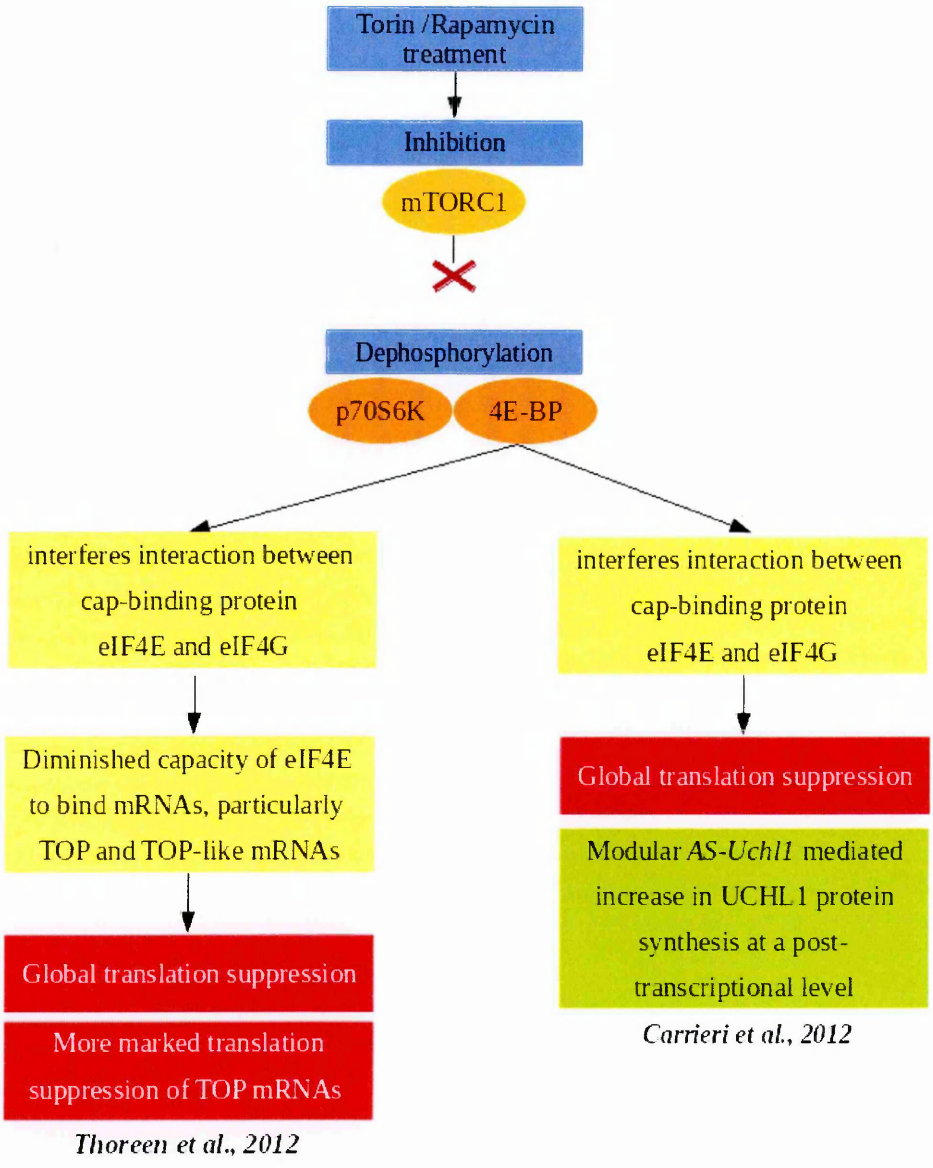


Figure 5.4 | Translation-control models involving mTORC1 inhibition. The above chart is a shematic representation of the translation control model proposed by *Thoreen et al., 2012*, (left) that involves TOP mRNAs and the post-transcriptional protein upregulatory activity of *AS-Uchl1* upon mTORC1 inhibition, experimental demonstrated by *Carrieri et al., 2012 (right).*

5.2. Materials and Methods

5.2.1. Functional enrichment analysis

To identify the influence of SINE orientation in ASlncRNA over the functional associations of overlapping sense coding genes, I performed a simple statistical proportion test using the *functional enrichment analysis module* of the pipeline. The proportion test is performed using the *prop.test()* function in R, where the proportion of the genes annotated for specific CC among each test gene groups (described in *figure 5.1)* are compared against the proportion of genes in control gene group, represented by the sense coding genes overlapping to ASlncRNAs without any SINE repeats. Here, the *prop.test()* function calculates the chi-squared statistic to test the null hypothesis according to which the proportion of genes annotated for specific CC is same between the test and the control set of genes. The alternative hypothesis is that the proportion of test genes annotated with specific CC is greater than that of the proportion of control gene group. Given that the multiple comparisons are performed in this analysis there is a need for the correction of the obtained p-values to narrow down the chances of false discoveries. For this, I implemented the FDR method based p-value adjustment, using *p.adjust()* function in R. In order not be very conservative, I performed the p-adjustment step for only those test gene sets that had a minimum of 15 annotated genes for a specific GO term. Finally, based on the adjusted p-values *(<= 0.05)*, the over-represented GO terms are selected for the representation into a comparative histograms showing the percentage of annotated genes for different test gene groups used in the analysis for further interpretations.

5.2.2. Analysis of translation efficiency in stress using the previously published data

To analyze the behavior of sense coding genes during normal and cellular stress, I decided to make use of the microarray data produced by *Giannakakis et al*, who investigated the levels of RNA associated to different polysome fractions in human MRC5 cell lysates during normal and oxidative stress conditions (Giannakakis et al., 2015). The polysome fractions in this study were classified into three pools based on the number of ribosomes found attached to the RNA molecules. If the RNA were identified to carry 5 or more ribosomes they were classified as high-translating (HT). Similarly, the RNA molecules with 2-4 and <2 ribosomes were classified as low-translating (LT) and not-translating (NT) fractions respectively. The custom designed microarray used by *Giannakakis et al* was capable of quantifying 22,001 lncRNAs (as per GENCODE annotations) and 17,535 randomly selected coding genes. To begin, I firstly intersected my list of sense coding genes *(described in figure 5.1)* and their respective ASlncRNA partners with the microarrray data, thereby extracting their relative RNA level in normal and stress conditions. Using this data-set, I investigated the shift of ASlncRNA with exclusively inverted or direct SINE repeats and their sense coding partners from either LT or NT to HT polysome fractions during normal and stress conditions. The main motive behind this investigation was to determine if the sense coding genes overlapping to specific ASlncRNA group show a shift in RNA levels from low to high translating polysome fraction in response to stress, which is a typical behavior expected from a coding gene that is under control of an ASlncRNA similar to *AS-Uchl1* or SINEUPs. For the ease in interpretation of RNA level changes in polysome fractions for different gene groups, I represented the data into comparative boxplots.

Further, I separately compared the relative RNA level ratio-of-mean (ROM) in stress over control for all ASlncRNA containing inverted and direct SINE repeats against noASlncRNAs in HT, LT and NT fractions. The motive behind this was to determine if the ASlncRNAs containing SINE repeats in specific orientation show a higher RNA levels specifically during stress, which is a characteristic similar to *AS-Uchl1* and other functional SINEUP. For this, I performed a randomization analysis by generating 1000 random samples from a total of 17948 noASlncRNA transcripts, where the sample size *n* were 305 (total no. of ASlncRNAs with inverted SINE) and 151 (total no. of ASlncRNAs with direct SINE) receptively.

I chose to perform a Z-test for this comparison because the sample size of the samples under comparison are large ($n > 30$), which makes the Z-test an appropriate test statistic to be used in this case (Ghasemi & Zahediasl, 2012). The Z-test builds upon the Z-score, which here is a measure of how many standard deviations below or above, the RNA level ROM of ASlncRNAs is, from that of the mean of ROM of noASlncRNAs. The formula used for Z-test is as follows: $\mathbf{z = (X - \mu) / \sigma}$, where z is the Z-score, X is the ROM for ASlncRNAs (stress/control), μ is the mean of the ROM of 1000 noASlncRNA random samples (stress/control) and finally σ is the observed standard deviation between the ROM of noASlncRNA random samples. The obtained Z-score is placed in the normal distribution to determine whether or not to reject the null hypothesis according to which the RNA level ROM of ASlncRNA and noASlncRNAs are the same. The Z-score is also used to calculate p-values using the *pnorm()* function in R, considering the two sided test as shown here, **p-value = 2 * pnorm(-abs(z))** *(the details of this step is discussed in section 2.2.4.3)*. Finally, I highlighted the ROM for ASlncRNAs containing exclusively inverted SINE and ASlncRNA containing

exclusively direct SINE repeats corresponding to HT, LT and NT fractions in the distribution of ROMs for 1000 random samples of noASlncRNA for further interpretations.

5.2.3. Analysis of the association of sense coding gene groups with mTORC1 signaling pathway

To test the hypothesis stated in section 5.1.1, I decided to use a comprehensive set of 1645 human TOP gene catalog collected by *Yamashita et al. 2008*, who used a position specific matrix (PSM) search for the TOP motifs that are usually defined to start with a "C" residue after the 5' cap-structure followed by 4–14 uninterrupted pyrimidine residues (Yamashita et al., 2008, Hamilton et al., 2006). I firstly overlapped my list of sense coding gene categories *(described in figure 5.1)* with the human TOP gene catalog, and analyzed the proportion of sense coding genes that contain a TOP motif. The main motive behind this analysis was to check, whether the TOP motifs are under-represented among sense coding genes overlapping to ASlncRNA containing inverted SINE repeats *(a test gene group)*, in contrast to coding genes overlapping to ASlncRNA without any SINE repeats *(control gene group)*. The reason for comparing these two gene groups is that the ASlncRNAs containing inverted SINE repeats resemble the most to *AS-Uchl1*, in comparison to rest of the gene categories described in *figure 5.1*. Hence, if they also share a similar functionality then the translation regulation of their respective overlapping coding genes should remain independent of the TOP motifs as stated in the hypothesis *(section 5.1.1)*. Also the coding genes overlapping to ASlncRNAs without any SINE are the best suited control set because they are least similar to the *AS-Uchl1*, as they lack the SINE elements which is the effector domain of the *AS-Uchl1*.

To compare of the proportions, I performed a simple two-sample proportion test using the the *prop.test()* function in R, to test the null hypothesis according to which the proportion of test genes sample with a 5'TOP motif is same as that of the proportion of control gene sample. For the sake of completeness, I also included rest of the other test gene groups in the analysis that contained a SINE repeat (shown in figure 5.1). Finally, I represented the percentage of genes containing a 5' TOP motif as comparative bar plots, corresponding to the test and control genes groups and highlight the derived p-values on the top each bar for further interpretation.

5.3. Results and discussions

5.3.1. Functional enrichment analysis considering SINE orientations in ASlncRNAs

The functional enrichment analysis revealed, the sense coding genes with ASlncRNA containing exclusively inverted SINE and exclusively direct SINE repeats are significantly enriched for mitochondrial localization *(GO term- GO:0005739)* in human as well as mouse *(adjusted p-value <= 0.05)* in contrast to sense coding genes with ASncRNAs containing no SINE repeats *(Figure 5.5)*. The result suggests that the sense coding genes overlapping to ASlncRNA that contain a SINE element are generally enriched for mitochondria specific annotations. And the orientation of SINE repeats within ASlncRNA do not necessarily contribute to a separate functional associations for the corresponding sense coding genes. The annotated gene names for human and mouse are shown in table *5.1* and *5.2* respectively

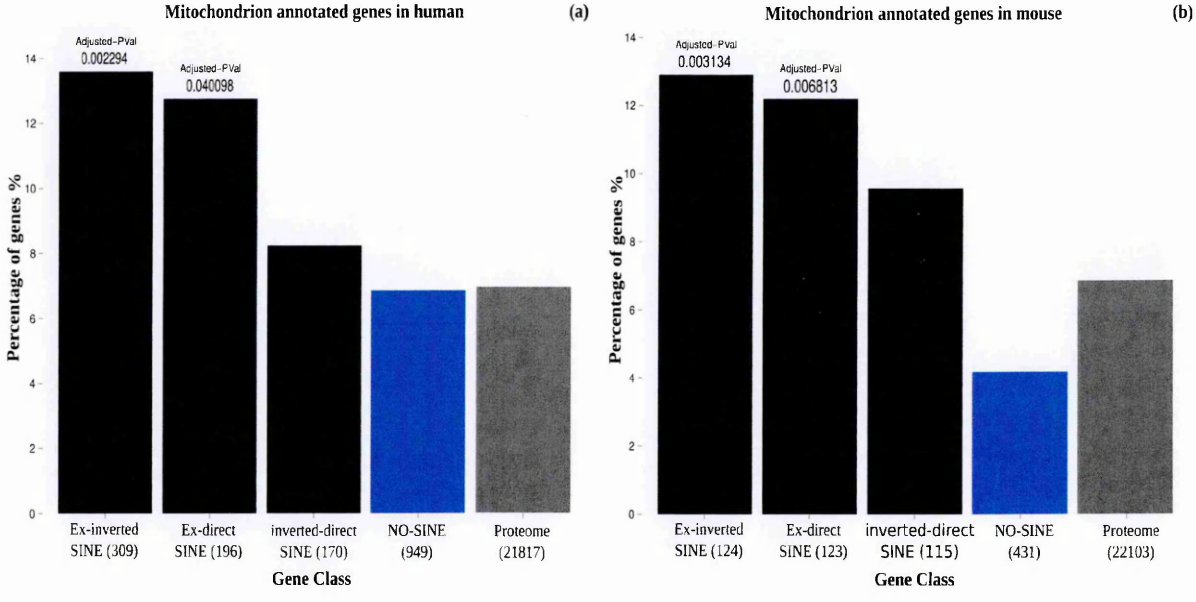


Figure 5.5 | Percentage of annotated genes. The charts represent the percentage of genes annotated for mitochondrion *(y-axis)* for each gene class (*x-axis)* for human (a) and mouse (b). The gene classes in *x-axis* are explained in figure 5.1.

Here, it is intriguing to observe that in *both* human and mouse, the sense coding genes that overlap to an ASlncRNA carrying SINE element are significantly enriched for mitochondrial localization. This suggests for a common functionality of ASlncRNAs containing SINE repeats across two vertebrate species. At the same time it can also be seen that none of the homologues genes in human and mouse that are annotated for mitochondrion, share an analogues SINE orientation (ex-inverted and ex-direct) property in their respective ASlncRNAs *(Table 5.3)*. This implies that although the ASlncRNAs containing a SINE element *(irrespective of its orientation)* might have an analogous functional implications on their respective sense coding genes in human and mouse, they do not necessarily share an analogues SINE orientation property at the same time.

Altogether, based on these results, one could hypothesize that the protein coding genes containing SINE repeats are essential for mitochondrial functions in a cell. And the embedded SINE elements in their overlapping ASlncRNA partner could act as the effector domain in a similar fashion to that of *AS-Uchl1* explained by *Carrieri et al*. However, the functional *AS-Uchl1* and SINEUPs particularly require inverted SINEs as the effector domain. To investigate the importance of SINE orientations within ASlncRNAs and its possible functional activity during stress, it is important to determine how their corresponding sense coding genes would behave in normal and cellular stress conditions. For this, I decided to use the data generated by *Giannakakis et al., 2015,* from a polysome fractionation experiment in human MRC5 cell lysates during cellular control and oxidative stress conditions and analyze the translation efficiency of the genes with ASlncRNA carrying inverted and direct SINE in response to stress.

Gene name	Ensembl gene id	SINE orientation	Gene name	Ensembl gene id	SINE orientation
AGK	ENSG00000006530	ex-inverted	SLC25A24	ENSG00000085491	ex-direct
TTC19	ENSG00000011295	ex-inverted	L2HGDH	ENSG00000087299	ex-direct
BCAT1	ENSG00000060982	ex-inverted	PYCARD	ENSG00000103490	ex-direct
DGAT2	ENSG00000062282	ex-inverted	SSBP1	ENSG00000106028	ex-direct
CS	ENSG00000062485	ex-inverted	NNT	ENSG00000112992	ex-direct
IDH3G	ENSG00000067829	ex-inverted	MTFR1L	ENSG00000117640	ex-direct
STARD7	ENSG00000084090	ex-inverted	AKAP1	ENSG00000121057	ex-direct
NDUFB2	ENSG00000090266	ex-inverted	GOT2	ENSG00000125166	ex-direct
GARS	ENSG00000106105	ex-inverted	PANK2	ENSG00000125779	ex-direct
MRPL51	ENSG00000111639	ex-inverted	CYP1B1	ENSG00000138061	ex-direct
WARS2	ENSG00000116874	ex-inverted	CLTC	ENSG00000141367	ex-direct
ROMO1	ENSG00000125995	ex-inverted	FEZ1	ENSG00000149557	ex-direct
TIMM17B	ENSG00000126768	ex-inverted	DHRS4	ENSG00000157326	ex-direct
OSGEPL1	ENSG00000128694	ex-inverted	HLCS	ENSG00000159267	ex-direct
BCL2L2	ENSG00000129473	ex-inverted	GFM2	ENSG00000164347	ex-direct
IMMT	ENSG00000132305	ex-inverted	GHITM	ENSG00000165678	ex-direct
DAP3	ENSG00000132676	ex-inverted	TPP1	ENSG00000166340	ex-direct
PEMT	ENSG00000133027	ex-inverted	GATM	ENSG00000171766	ex-direct
CMPK2	ENSG00000134326	ex-inverted	BSG	ENSG00000172270	ex-direct
USP30	ENSG00000135093	ex-inverted	SPRYD4	ENSG00000176422	ex-direct
TMEM8B	ENSG00000137103	ex-inverted	D2HGDH	ENSG00000180902	ex-direct
IDH1	ENSG00000138413	ex-inverted	ERCC6L2	ENSG00000182150	ex-direct
RAP1GDS1	ENSG00000138698	ex-inverted	CARKD	ENSG00000213995	ex-direct
C12orf10	ENSG00000139637	ex-inverted	CAPN1	ENSG00000014216	inverted-direct
ALDH1L1	ENSG00000144908	ex-inverted	PITRM1	ENSG00000107959	inverted-direct
STAR	ENSG00000147465	ex-inverted	RHOT1	ENSG00000126858	inverted-direct
LRRK1	ENSG00000154237	ex-inverted	DUT	ENSG00000128951	inverted-direct
COQ7	ENSG00000167186	ex-inverted	CYP11A1	ENSG00000140459	inverted-direct
COA6	ENSG00000168275	ex-inverted	SNCA	ENSG00000145335	inverted-direct
MFF	ENSG00000168958	ex-inverted	SYBU	ENSG00000147642	inverted-direct
HARS	ENSG00000170445	ex-inverted	CDKN2A	ENSG00000147889	inverted-direct
RNASEH1	ENSG00000171865	ex-inverted	SLC16A1	ENSG00000155380	inverted-direct
TEFM	ENSG00000172171	ex-inverted	HK2	ENSG00000159399	inverted-direct
METAP1D	ENSG00000172878	ex-inverted	TK2	ENSG00000166548	inverted-direct
DHFRL1	ENSG00000178700	ex-inverted	UQCRFS1	ENSG00000169021	inverted-direct
ADO	ENSG00000181915	ex-inverted	TRIM39	ENSG00000204599	inverted-direct
TDRKH	ENSG00000182134	ex-inverted	ATP5O	ENSG00000241837	inverted-direct
GLRX5	ENSG00000182512	ex-inverted			
NDUFA6	ENSG00000184983	ex-inverted			
RAB11B	ENSG00000185236	ex-inverted			
TRIM31	ENSG00000204616	ex-inverted			
GNB2L1	ENSG00000204628	ex-inverted			
OXCT1	ENSG00000083720	ex-direct			

Table 5.1 | Human sense coding genes annotated for mitochondrion.

Gene name	Ensembl gene id	SINE orientation	Gene name	Ensembl gene id	SINE orientation
Taco1	ENSMUSG00000001983	ex-inverted	Bcat1	ENSMUSG00000030268	ex-direct
Tomm40	ENSMUSG00000002984	ex-inverted	Bcat2	ENSMUSG00000030826	ex-direct
Nars2	ENSMUSG00000018995	ex-inverted	Me1	ENSMUSG00000032418	ex-direct
Comtd1	ENSMUSG00000021773	ex-inverted	Brinp3	ENSMUSG00000035131	ex-direct
Mtpap	ENSMUSG00000024234	ex-inverted	Tigar	ENSMUSG00000038028	ex-direct
Cidea	ENSMUSG00000024526	ex-inverted	Bdh1	ENSMUSG00000046598	ex-direct
Mgme1	ENSMUSG00000027424	ex-inverted	Pfdn4	ENSMUSG00000052033	ex-direct
Pus1	ENSMUSG00000029507	ex-inverted	Mdh1	ENSMUSG00000020321	inverted-direct
Eln	ENSMUSG00000029675	ex-inverted	Oxct1	ENSMUSG00000022186	inverted-direct
Alas1	ENSMUSG00000032786	ex-inverted	Capn1	ENSMUSG00000024942	inverted-direct
Prr5l	ENSMUSG00000032841	ex-inverted	Apex-2	ENSMUSG00000025269	inverted-direct
Tomm6	ENSMUSG00000033475	ex-inverted	Mff	ENSMUSG00000026150	inverted-direct
Nudt19	ENSMUSG00000034875	ex-inverted	Mccc1	ENSMUSG00000027709	inverted-direct
Kif1bp	ENSMUSG00000036955	ex-inverted	Abcb1b	ENSMUSG00000028970	inverted-direct
Mrps26	ENSMUSG00000037740	ex-inverted	Aldh5a1	ENSMUSG00000035936	inverted-direct
1700123O20Rik	ENSMUSG00000040822	ex-inverted	Tmem11	ENSMUSG00000043284	inverted-direct
Coa3	ENSMUSG00000017188	ex-direct	2310061I04Rik	ENSMUSG00000050705	inverted-direct
Slc25a19	ENSMUSG00000020744	ex-direct	Rab11b	ENSMUSG00000077450	inverted-direct
Acox1	ENSMUSG00000020777	ex-direct			
Ripk1	ENSMUSG00000021408	ex-direct			
Fen1	ENSMUSG00000024742	ex-direct			
Ak3	ENSMUSG00000024782	ex-direct			
Acox3	ENSMUSG00000029098	ex-direct			
Mrpl53	ENSMUSG00000030037	ex-direct			

Table 5.2 | Mouse sense coding genes annotated for mitochondrion. Here, *ex-inverted* and *ex-direct* is used to represent ASlncRNA containing exclusively inverted and exclusively direct SINE repeats. Whereas, *inverted-direct* represents ASlncRNA containing both inverted and direct SINE repeat.

No.	Gene name	Human SINE orientation	Mouse SINE orientation	Analogues 3' domain	Mitochondria Annotation	Analogues 3' domain & Mitochondria Annotation
1	CAPN1	inverted-direct	inverted-direct	1	1	1
2	BCAT1	ex-inverted	ex-direct	0	1	0
3	SRCAP	ex-inverted	inverted-direct	0	0	0
4	OXCT1	ex-direct	inverted-direct	0	1	0
5	NKAIN4	ex-inverted	ex-direct	0	0	0
6	DDX59	ex-inverted	ex-inverted	1	0	0
7	MORF4L2	inverted-direct	inverted-direct	1	0	0
8	PAX8	inverted-direct	ex-inverted	0	0	0
9	UXT	ex-inverted	ex-inverted	1	0	0
10	MKLN1	ex-inverted	inverted-direct	0	0	0
11	BHLHE40	ex-inverted	inverted-direct	0	0	0
12	UCHL1	inverted-direct	inverted-direct	1	0	0
13	DEPTOR	inverted-direct	inverted-direct	1	0	0
14	MFF	ex-inverted	inverted-direct	0	1	0
15	PCBP1	inverted-direct	inverted-direct	1	0	0
16	EMX2	inverted-direct	inverted-direct	1	0	0
17	PDE3A	ex-inverted	ex-direct	0	0	0
18	DDN	inverted-direct	ex-direct	0	0	0
19	UNC5C	ex-direct	ex-direct	1	0	0
20	RAB11B	ex-inverted	inverted-direct	0	1	0
21	DIO2	ex-direct	ex-inverted	0	0	0
22	COLCA2	ex-inverted	ex-inverted	1	0	0

Table 5.3 | Homologous sense coding genes in human and mouse. The above table contains the list of all sense coding genes with head-to-head ASlncRNA overlap that are homologous in human and mouse. Column 3, 4 from left represents the SINE orientation in the their respective ASlncRNAs. Column 4 is marked with a value 1, if the SINE orientation in ASlncRNA is analogous in human and mouse. Similarity, column 5 is marked with value 1, if the genes in human and mouse are annotated to mitochondria specific gene ontology. Finally, the last column is marked with value 1, if the genes are annotated to mitochondrion and their respective ASlncRNA have an analogous SINE orientation in human and mouse

5.3.2. Analysis to identify the translation efficiency of sense coding genes in stress

To test the translation efficiency of the genes with ASlncRNA carrying inverted and direct SINE repeat in response to stress, I overlapped my list of genes with the microarrray data generated from *Giannakakis et al* and investigated the shift of ASlncRNA and their corresponding sense coding genes from LT/NT polysome fractions to HT fractions in stress (as discussed in 5.2.2). The result of the overlap analysis revealed that neither group of ASlncRNA carrying inverted or direct SINE and their corresponding sense coding genes show any significant differential polysome loading in stress with respect to normal conditions *(Figure 5.6)*. Similar observations were also accounted in case of ASlncRNAs overlapping to the sense coding genes that are specifically annotated for mitochondrion localization *(Table 5.1, 5.2; Figure 5.7)*.

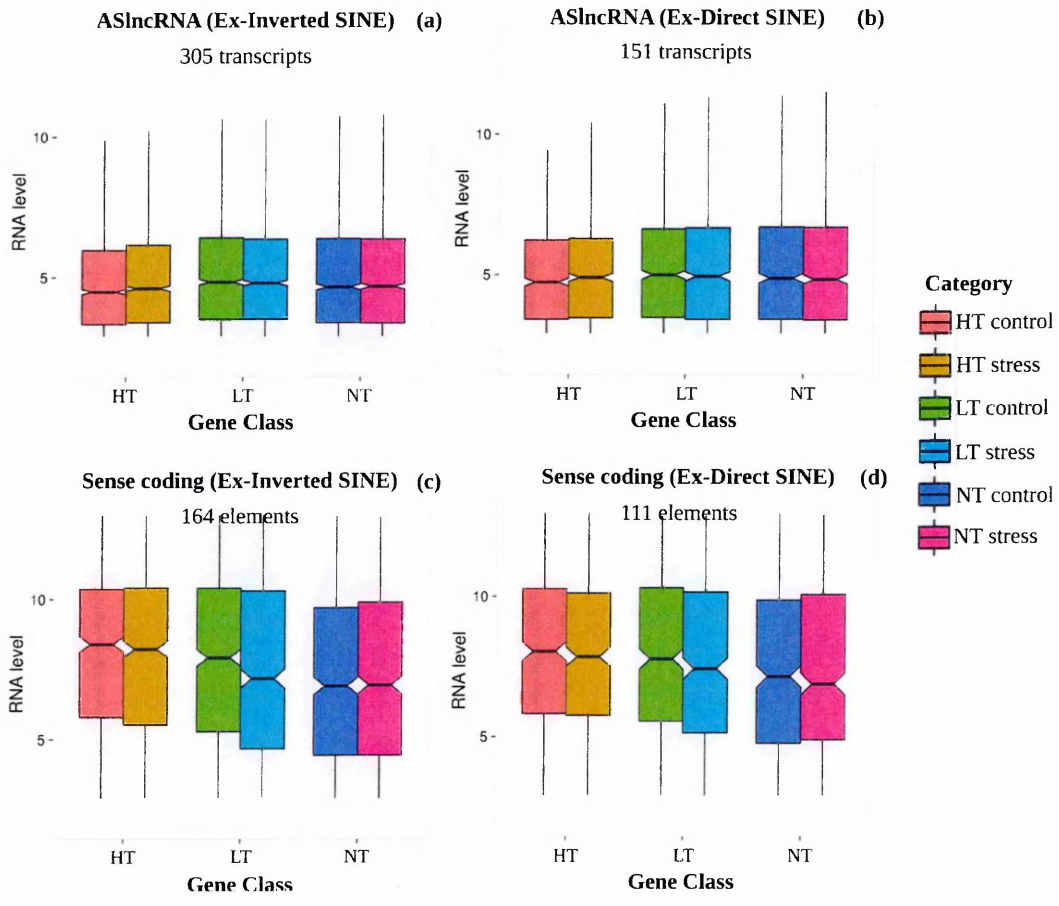


Figure 5.6 | Translational switch of transcripts in response to stress. In the above charts, *x-axis* represents the three polysome fraction classified as HT (high-translating), LT (low translating) and NT (non-translating) in control and stress conditions, *y-axis* represents the relative RNA levels for the (a) ASlncRNAs with Ex-inverted-SINE; (b) AslncRNA with Ex-direct-SINE; (c) Sense coding elements with ASlncRNA containing Ex-inverted-SINE; (d) coding elements with AslncRNA containing Ex-direct-SINE. Here, the transcripts do not show a significant shift from low to high translating polysome fractions during stress in comparison to the control, also ASlncRNA containing exclusive inverted or direct SINE repeats do not show any difference.

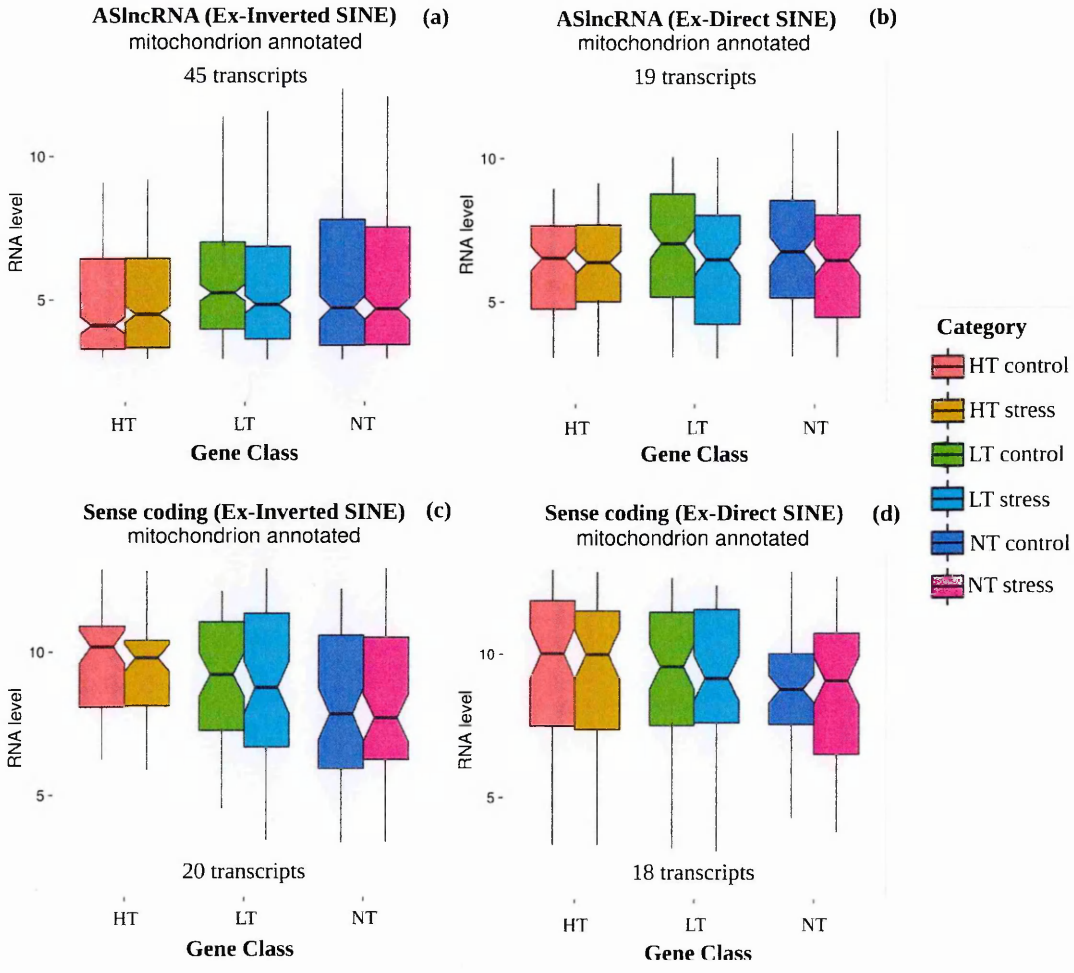


Figure 5.7 | Translational switch of transcripts in response to stress (mitochondrion annotated sense coding genes and their corresponding ASlncRNA transcripts) (a) ASlncRNAs with Ex-inverted-SINE; (b) ASlncRNA with Ex-direct-SINE; (c) Sense coding elements with ASlncRNA containing Ex-inverted-SINE (d) coding elements with ASlncRNA containing Ex-direct-SINE. Here, ASlncRNA and sense coding genes do not show any significant shift of RNA levels from low to high polysome fraction during stress in contrast to normal cell conditions.

To be the good candidates for SINEUP like activity, the sense coding genes overlapping to ASlncRNA containing inverted SINEs are expected to show the stress related shift of from low to high translating polysome fractions similar to what has been reported for *AS-Uchl1* and functional SINEUPs. However, no significant change in the RNA levels were observed for either of the sense coding gene categories that overlap to ASlcnRNAs carrying inverted or direct SINEs. This suggests, the list ASlncRNAs containing inverted SINE repeats do not act as the good candidates for SINEUP like activity at least based on the observation made using the microarray data from *Giannakakis et al.,'s* polysome fractionation experiment in human MRC5 cell lysates. However, at the same time we also can not rule out the fact that the RNA level data corresponding to protein coding-genes were incomplete, as only 17,535 randomly selected coding gene were quantified by the designed custom microarray used by *Giannakakis et al.* This could be one of the reasons for the observed no differences in RNA levels for different polysome fractions during cellular stress vs normal conditions. To confirm this there is a requirement of performing a similar polysome fractionation experiment but this time targeting all the genes in the gene list described in *figure 5.1*, this would help to shed more light on the contribution of SINE content/orientation within ASlncRNA over the behavior of their corresponding sense coding genes in stress response.

Nevertheless, to extract meaningful information from the data for ASlncRNAs, I decided to compare the RNA level ratio of means (ROM) in stress over control for ASlncRNAs containing inverted and direct SINE repeats against the ROM in stress over control for noASlncRNAs in HT, LT and NT fractions. Such a comparison would reveal if ASlncRNAs containing SINE repeats in specific orientation, show a different stress responsive change in RNA levels, in the light of, their differences against noASlncRNAs. I performed this

comparison based on a randomization analysis as discussed in section ***5.2.2*** *(second paragraph).*

Interestingly, the result of this comparative randomization analysis revealed that the ASlncRNAs containing a SINE repeat *(irrespective of its orientations)* show significantly higher RNA levels in response to stress particularly in HT fraction, in contrast to noASlncRNAs *(Figure 5.8 a, b).* This behavior resembles to the characteristic of *AS-Uchl1* in stress, however no evidence for the stress responsive increase in polysome loading of their respective sense coding genes were seen in the comparative analysis discussed above *(Figure 5.6, 5.7; c,d).* Altogether, based on this analysis it can be concluded that the ASlncRNAs containing SINE repeats are the key RNA molecules that are active during stress and additional similar experimental validation targeting all the genes in the gene list described in *figure 5.1* and their respective ASlncRNAs could help to better understand the stress responsive changes of the sense coding genes that overlap to ASlncRNA containing SINE repeats.

the *x-axis*. The ratio of mean for ASlncRNAs containing (a, c, e) exclusively inverted SINE and ASlncRNA containing exclusive direct SINE (b, d, f) repeats corresponding to HT, LT and NT fractions respectively are highlighted in red colored dotted line. The Z-scores for ASlncRNAs containing inverted SINE repeats in different RNA fraction are as follows – *HT: 6.230996, LT: -0.5544477, NT: -0.03454891*. Similarly, the Z-scores for ASlncRNAs containing direct SINE repeats are - *HT: 3.543415, LT: -0.7127415, NT: -1.451755*. Here, we can observe that the ASlncRNA particularly in HT RNA fractions, containing either inverted or direct SINE repeats show a significant higher ratio of mean for RNA levels in stress over control than noASlncRNAs.

5.3.3. Analysis of 5'-TOP motif enrichment among sense coding genes

The results of the analysis aiming to test the *"Deprived 5' TOP motif hypothesis"* stated in section 5.5.1 revealed that the genes overlapping to ASlncRNA with exclusively direct SINEs are significantly enriched for TOP motif representation, whereas there is no significant difference observed in the representation of TOP motifs between the genes overlapping to ASlncRNA containing exclusive inverted SINEs (Ex-inverted) and the genes overlapping to ASlncRNAs with no SINE (NO-SINE) *(Figure 5.9)*. This suggests that the genes with ASlncRNAs containing exclusive inverted SINE are not likely to rely on mTORC1 mediated translation control that involves TOP motifs, instead they could be the good candidates to act similar to *AS-Uchl1*, because they share a similar modular organization and an inverted SINE element. At the same time, it is also interesting to observe that the sense coding genes with ASlncRNA containing direct SINE repeats show a significant over-representation of TOP motifs, to understand why, there is a need of further exploration.

5.4. Conclusions

The results of the analysis described in this chapter suggests that the sense coding genes with ASlncRNA containing a SINE repeat *(irrespective of its orientation)* are generically enriched for mitochondrion specific annotation in human and mouse. This suggests for a common functionality of ASlncRNAs containing SINE repeats across two vertebrate species. However, only ~3% of all human sense coding genes in head-to-head overlap with ASlncRNA containing a SINE have an annotated homologous mouse sense coding gene in similar head-to-head overlap with an ASlncRNA *(Table 5.3)*. Also among these homologous set of genes, the one that are also commonly annotated for mitochondrion, do not share an analogous SINE orientation property. This implies, that although the ASlncRNAs containing a SINE element have an analogous functional implications on their respective sense coding genes in human and mouse, they do not necessarily be homologous to each other or share an analogues SINE orientation property.

Further, ASlncRNAs carrying inverted or direct SINE repeats and their corresponding overlapping protein coding genes did not show a significant shift of RNA levels from low to high polysome fractions in response to stress. However interestingly, the randomization analysis considering noASlncRNAs revealed that the ASlncRNAs containing a SINE elements show significantly higher RNA levels in response to stress, particularly in the high translating fractions. This suggests that the ASlncRNAs containing SINE repeats are active in response to stress and could be involved in translational up-regulation, because they demonstrate significant increase in expression during during stress which is a similar characteristic as that *AS-Uchl1* However, there is a need of further experiments to determine the shift of RNA levels within different RNA fractions of the cell, because the microarray data used in the analysis

were corresponding to incomplete list of coding genes. And coding genes are the ones which are translated in an S/AS pair, therefore, determining the polysome loading of coding mRNA corresponding to ASlncRNA containing specific SINE repeats, and their relative shift between RNA fractions in response to stress would be more informative and helpful to conclude the role of ASlncRNAs..

Finally, the analysis to examine over representation of 5' TOP motif revealed, coding genes overlapping to ASlncRNA containing exclusively inverted SINE do not show any significant difference for the TOP motif representation in comparison to coding genes with ASlncRNA containing no SINE. This lack of over representation of 5' TOP motifs and similar modular organization of their ASlncRNA to that of *AS-Uchl1* in terms of SINE orientation, makes them good candidates to be further tested for SINEUP like activity. At the same time, the significant over representation of TOP motifs in sense coding gene overlapping to ASlncRNAs containing direct SINE, suggests these ASlncRNAs could be involved in regulating the sense coding genes that are involved in mTORC1 signaling pathway. Another point to be noted is that the human TOP gene catalog used in this analysis is produced by *Yamashita et al., in 2008.* Although it is the largest available catalog, it does not represent the most updated and exhaustive list of TOP genes. Hence, further identification of TOP motifs using similar techniques used by *Yamashita et al.* could help to expand the TOP gene catalog, as well as give strength to the TOP enrichment analysis performed in this chapter.

Chapter 6

General conclusions, discussions and future perspectives

The aim of this thesis was to explore and define the functional association between ASlncRNAs and TEs in the light of functional activity of the modular *AS-Uchl1* described by *Carrieri et al., 2012*. Such an exploration requires the following questions to be answered -

- Are ASlncRNAs enriched for SINE repeats in contrast to rest of the lncRNAs?
- How do the 5' overlapping domain and the 3' effector domain of ASlncRNA, influence the functional activity of the overlapping sense coding genes?
- Could ASlncRNAs with similar modular domains as that of *AS-Uchl1,* exert similar functional activity?

In this chapter, I discuss to what extent the above three objectives have been achieved and what are the general conclusions and the biological implications of the observations made in my study.

6.1. SINE TEs are the major contributors to the diversification of mammalian lncRNAs

Recently published studies have characterized the TEs content of lncRNAs and have shown that TEs are non-randomly distributed across lncRNAs. TEs are also known to cover a substantial portion of the total lncRNA sequences in human, mouse and other vertebrates (Kelley & Rinn, 2012; Kapusta et al., 2013). Based on the observations made in my study, I report that the SINE TEs are significantly enriched among ASlncRNAs sequences in contrast to other lncRNAs.

The SINE TEs are known to be present in abundance within the genomes of human and mouse in comparison to non-mammalian vertebrates and other invertebrates. With the split of primate-rodent lineages, SINEs have experienced lineage specifics TEs dynamics. As a consequence, today we observe distinct SINE types in human and mouse genomes (Sela et al., 2010; Silva et al., 2003; Smit & Riggs, 1995). This is also reflected in context to the human and mouse specific ASlncRNAs analyzed in my study. The SINE subfamilies/elements that are identified to be significantly enriched among ASlncRNAs of human and mouse, belonged particularly to the most ancient SINE class that are know to have diverged from a common origin prior to the split of primate-rodent lineages *(described in chapter 3).* This might imply that, although the SINE elements have taken different evolutionary route after the split, they have commonly contributed towards the evolution of ASlncRNAs in human and mouse.

6.2. Functional influence of the 5' binding domain remains elusive

The 5' binding domain of *AS-Uchl1* and synthetic SINEUPs is a complementary sequence centering the ATG of the target sense mRNA. It is said to be involved in providing the specificity for their targeted binding to mRNAs (Carrieri et al., 2012, Zucchelli et al., 2015a; Yao et al., 2015). However, the underlying mechanism through which the translation up-regulation of the sense *Uchl1* mRNA occurs is not known. As a consequence, the influence of ATG overlap over the translation initiation is also not understood. Based on the observations made in my study, I report that the ATG overlap of sense coding gene by an ASlncRNAs *(that are expected to act as AS-Uchl1)*, is unlikely to be involved in modulating the translation of sense coding gene at-least by TIS switch, in a way that the resultant protein attains a mitochondria specific localization signal. This is clearly seen at least in the case of mouse, where the ATG overlapped sense coding genes were found to be significantly enriched for the

mitochondria specific GO annotation *(discussed in chapter 4)*, however the TIS switch hypothesis *(described in figure 6.1)* did not hold true for these set of genes as well.

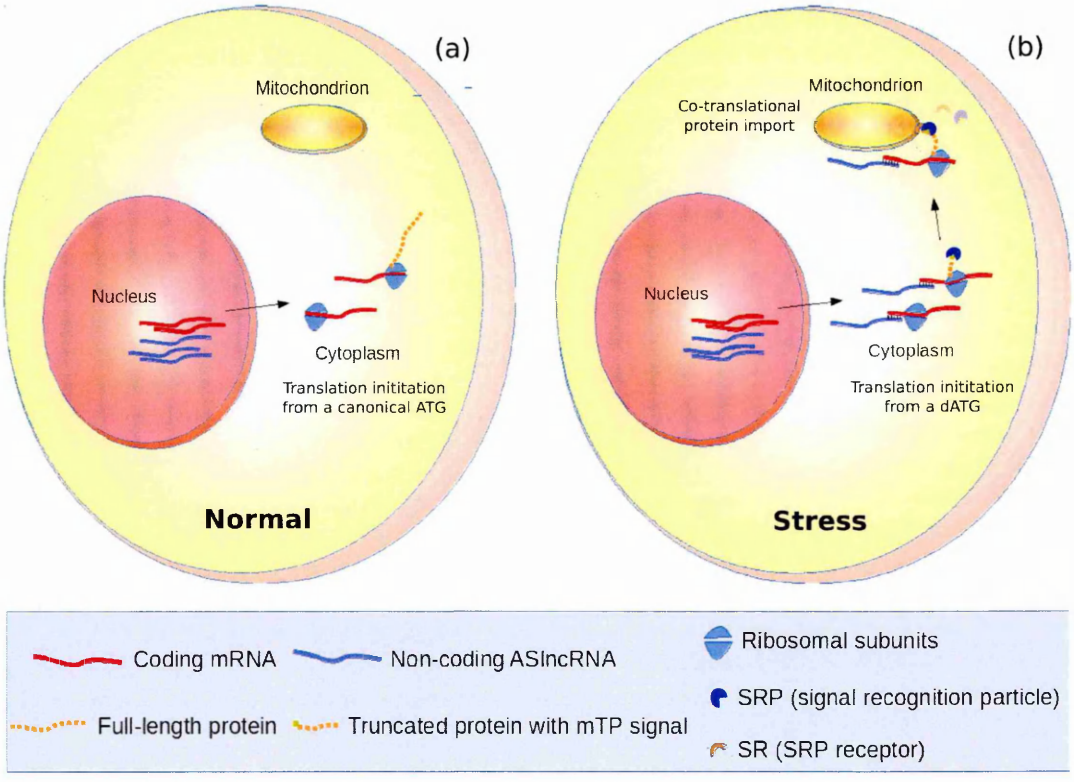


Figure 6.1 | TIS switch hypothesis *(modified in context to mitochondrion specific functional enrichment of sense coding genes with ATG overlap)*. The above cartoon represents translation of sense coding genes in (a) normal and (b) stressed conditions. During the normal condition, the mRNA corresponding to a sense coding gene gets translated in cytoplasm. Here, the translation initiation occurs from the canonical ATG present at the 5' end of the coding mRNA, yielding a full length protein sequence. However, in case of cellular stress the ASlncRNAs (that are expected to act as *AS-Uchl1*) are shuttled from nucleus to cytoplasm,

where they bind to mRNAs in a target-specific manner, as explained for *AS-Uchl1*. Due to the overlap of ASlncRNA the canonical ATG is blocked. Therefore, the translation initiation occurs from a non-canonical downstream ATG, that yields a truncated form of the protein with a different N terminus sequence that may contains an mTP signal peptide (mitochondrion specific target peptide). The translocation of such proteins into the mitochondrion occurs though the co-translational protein import pathway. This involves the signal recognition particle (SRP) to deliver truncated form of the protein to mitochondrion while they are still being synthesized by ribosomes (Saraogi & Shan, 2011). It is important to note that there are several other protein targeting pathways (Lithgow, 2000; Saraogi & Shan, 2011), however I have chosen the co-translational targeting of proteins by the signal recognition particle (SRP) to explain the TIS switch hypothesis, because it is one of the most extensively studied protein targeting pathways with excellent model system for in-depth mechanistic dissections to uncover the molecular basis of cellular protein localization (Mukhopadhyay, Ni, & Weiner, 2004).

6.3. ASlncRNAs containing SINEs are generally associated with nuclear genes encoding mitochondrial proteins

The *AS-Uchl1* and synthetic SINEUPs are described to contain an *inverted* SINE repeat which is the 3' effector domain and is required to exert the post-transcriptional protein up-regulation of the overlapping sense coding genes (Carrieri et al., 2012, Zucchelli et al., 2015a). Although the underlying mechanism to explain how the inverted SINE is involved in the protein up-regulation activity is not known. Also it is not understood, how the inverted SINE is different from direct SINE in influencing the functional activity of *AS-Uchl1* and SINEUPs. Based on the observations made in my study, I report that the orientation of the embedded SINE repeats in ASlncRNAs do not necessarily influence the functional association of the overlapping sense coding genes. In fact, the ASlncRNAs embedded with a SINE element (irrespective of SINE orientation) are associated with the sense coding genes that are significantly enriched for mitochondiron specific GO annotations in both human and mouse *(discussed in chapter 5)*. However, none of the mitochondrion associated genes in human and mouse shared an identical S/AS pair configuration (overlap type and and SINE orientation).

This implies that the ASlncRNA containing a SINE element *(irrespective of its orientation)* might have analogous functional implications in human and mouse. The observation made here also supports the notion that the SINE elements have played an important role, not only in the evolution of ASlnRNAs but also in the functional diversification in lncRNAs. Finally in my study, I did not find any evidence for the stress induced translational up-regulation activity of ASlncRNA embedded with inverted or direct SINE, over the overlapping sense coding genes *(discussed in chapter 5)*.

6.4. Coding genes overlapping to ASlncRNAs containing inverted SINEs are less likely to undergo 5' TOP motif involved mTORC1 translation-control

Uchl1 mRNA is shown to defy the mTORC1 translation-control, where the inhibition of mTORC1 suppress the global translation rate, a marked increase in translation is seen for *Uchl1* that is mediated by the overlapping modular *AS-Uchl1* (Carrieri et al., 2012*)*. In another study it has been suggested that the mRNAs that are regulated by mTORC1 contains a 5' terminal oligopyrimidine (TOP) motif, and such mRNAs show a more marked translation suppression upon the mTORC1 inhibition in comparison to the slightly impaired global translation (Thoreen et al., 2012). In my study, I tested an hypothesis that all the coding genes that overlap to an ASlncRNA containing inverted SINEs should be deprived of 5' TOP motifs, because they resemble to the S/AS pair of *Uchl1* and *AS-Uchl1* in terms of the overlap type and the modular organization of antisense, therefore are expected to behave similarly upon the mTORC1 inhibition *(discussed in chapter 5).*

Based on my investigation, I report that indeed only a small fraction of coding genes with ASlncRNA carrying inverted SINE are know to have a 5' TOP motif which is not significantly different from TOP containing fraction of genes that overlap to ASlncRNA without any SINE elements. This implies that the sense coding genes overlapping to ASlncRNA containing inverted SINEs do not represent the subset of TOP mRNAs that are regulated by mTORC1. At the same time I also report that a significantly higher fraction of coding genes with ASlncRNAs embedded with direct SINEs, contain a TOP motif. This implies a differential influence of SINE orientation in ASlncRNAs upon overlapping coding genes. However to confirm this, there is a need of further exploration and identification of a more complete catalog of TOP genes which basically relies on the accurate identification of the TSSs

(Yamashita et al., 2008). The large-scale genome-wide accurate identification of the TSS and hence the TOP motifs, could be further improvised by considering different approaches for example the HeliScopeCAGE technique coupled with different motif discovery methods employed by *Eliseeva et al., 2013* could help to build a comprehensive list of TOP or TOP like genes.

6.5. Concluding Remarks

The work presented in this Ph.D. thesis provides broader insights on natural ASlncRNA that are similar to *AS-Uchl1* in terms of their modular organization i.e, 5' end specific binding domain and 3' end specific effector domain. The novel bioinformatic approaches presented in this thesis illustrates different characteristics of the modular nature of ASlncRNAs and their influence over the functional activity of sense coding genes. Although, no strong evidence were accounted to explain if the natural ASlncRNAs could function as *AS-Uchl1*, the work presented here have highlighted important aspects of natural ASlncRNAs such as, the SINE specific sequence coverage enrichment in contrast to noASlncRNAs, SINE associated mitochondria specific functional enrichment of their overlapping sense coding genes and that the sense coding genes associated to ASlncRNA containing inverted SINE repeats, do not represent TOP mRNAs that undergo mTORC1 dependent translation suppression. Additionally, the TIS switch the hypothesis testing discussed in this thesis indicated that the ATG overlap of sense coding genes by the 5' binding domain of ASlncRNA is unlikely to be involved in modulating the translation of sense coding gene at-least by TIS switch, in a way that the resultant protein attains a mitochondria specific localization signal. Altogether, the work presented in this thesis can be used as the directions for further experimental scrutinization and bioinformatic exploration.

6.6. Future perspectives

Based on the observations made in this Ph.D. thesis, following are the main future actions which would further improve the resolution of our current understanding of modular characteristics of ASlncRNAs, and reveal if the ASlncRNAs could act as *AS-Uchl1*

6.6.1. Experimental testing of TIS switch hypothesis

Mitochondrion specific annotation enrichment among sense coding genes is an important aspect associated to ASlncRNAs, because this has been accounted independently for both the ATG overlap and SINE repeat content characteristics of ASlncRNAs. Therefore, mitochondrion specific annotation enrichment could be used as the basis for the investigation of the modular nature of ASlncRNA. The bioinformatic approach for such an investigation has been explained as the testing of "TIS switch hypothesis" in my thesis *(Figure 6.1)*, which revealed no relation between the ATG overlap and TIS switch based translation and the change in protein sub-cellular localization. However this analysis was totally based on bioinformatic prediction of signal peptides, perhaps an experimental approach to test the TIS hypothesis could reveal better information. Following are two possible experimental procedures which could be used to test the TIS switch hypothesis -

- ***proximity-specific ribosome profiling:*** It is a technique used to measure translation at the mitochondrial surface in yeast. It involves *in vivo* biotinylation of Avi-tagged ribosomes that are in contact with a spatially localized biotin ligase, followed by affinity purification of biotinylated ribosomes and measure of translational activity by deep sequencing of ribosome-protected fragments (Williams, Jan, & Weissman, 2014).

This technique could be used to test the TIS switch hypothesis, where the synthetic constructs of the full-length and truncated mRNAs can be transfected into the separate cells, followed by the application of mitochondria proximity-specific ribosome profiling. If the TIS switch hypothesis holds true then a higher percentage of truncated mRNA would be quantified near to the mitochondrial proximity in comparison to the full-length mRNAs.

- ***The live-cell imaging:*** Is a technique for the direct visualization of the real-time transport of RNA molecules in the cell. There are several high-end visualization techniques available (Buxbaum, Haimovich, & Singer, 2015; Zepeda et al., 2013). The live-cell imaging technique can also be used for testing the TIS switch hypothesis, by monitoring the real-time translocation of full length and truncated mRNAs that are trasfected into separate cells.

6.6.2. Polysome fractionation experiment

The polysome fractionation experiment performed by *Giannakakis et al, 2015 (discussed in chapter 5)*, is a best suited experiment to quantify a stress responsive RNA level shift between the high and low translating RNA fraction within a cell for a set of genes (*described in 5.2.2*). However, the comparative analysis performed by me using this data, did not show any stress responsive significant shift of RNA levels between different polysome loading fractions for sense coding genes, which is a desired characteristic for a gene that is in overlap with *AS-Uchl1* like ASlncRNA. The reason for such an observation could be the fact that only a subset of coding genes were randomly selected to be quantified in the custom designed microarray used by *Giannakakis et al., 2015*. Hence, performing a similar polysome fractionation

experiment as explained by *Giannakakis et al,* followed by a custom microarray targeting to quantity my list of identified and characterized sense coding genes and their respective overlapping ASlncRNA partners would help to reveal better information regarding their stress responsive behavior in a cell.

6.6.3. Predicting RNA secondary structure and RNA-RNA interaction

In last few decades RNA secondary structure prediction has emerged as a key step to understand in-sillico identification of RNA-RNA interaction such as, the interactions between the candidate non-coding RNA and their targets (Meyer, 2008). As a consequence, a large compendium of RNA secondary structure prediction tool is available today (ink: https://en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software). Analyzing the interaction between the secondary structure of the S/AS pair of *Uchl1* mRNA and *AS-Uchl1* and similar such natural sense coding genes and their overlapping ASlncRNAs might greatly help to understand the yet unknown mechanism underlying the post-transcriptional translation up-regulation of *Uchl1* mRNA mediated by the *AS-Uchl1*. Such an analysis might also reveal a possible conserved secondary structures attained by the similar S/AS pair of transcripts and understand how inverted SINE elements could be involved in the translation process explained in case of *AS-Uchl1*. The secondary structure based RNA-RNA interaction might also greatly help to understand how different would be the interaction of ASlncRNA carrying an inverted SINE and ASncRNA carrying a direct SINE repeats with their respective sense coding mRNA overlapping partners.

References

Adams, M., Celniker, S., Holt, R., Evans, C., Gocayne, J., Amanatides, P., … Venter, J. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195. http://doi.org/10.1126/science.287.5461.2185

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Molecular Biology of the Cell. Garland Science. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21054/

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. http://doi.org/10.1016/S0022-2836(05)80360-2

Amaral, P. P., & Mattick, J. S. (2008). Noncoding RNA in development. *Mammalian Genome Official Journal of the International Mammalian Genome Society*, *19*(7-8), 454–492. http://doi.org/10.1007/s00335-008-9136-7

An, W., Han, J. S., Schrum, C. M., Maitra, A., & Boeke, J. D. (2009). Conditional Activation of a Single-Copy L1 Transgene in Mice by Cre, *46*(7), 373–383. http://doi.org/10.1002/dvg.20407.Conditional

Ard, R., Tong, P., & Allshire, R. C. (2014). Long non-coding RNA-mediated transcriptional interference of a permease gene confers drug tolerance in fission yeast. *Nature Communications*, *5*, 5576. http://doi.org/10.1038/ncomms6576

Arrial, R. T., Togawa, R. C., & Brigido, M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics*, *10*(1), 239. http://doi.org/10.1186/1471-2105-10-239

Ashburner, M., & Bergman, C. M. (2005). Drosophila melanogaster : A case study of a model genomic sequence and its consequences, 1661–1667. http://doi.org/10.1101/gr.3726705.15

Athanikar, J. N., Badge, R. M., & Moran, J. V. (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Research*, *32*(13), 3846–3855. http://doi.org/10.1093/nar/gkh698

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *Am. J. Hum. Genet, 73*, 823–834. http://doi.org/10.1086/378594

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. a, De Sapio, F., … Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature, 479*(7374), 534–7. http://doi.org/10.1038/nature10531

Barrangou, R., Birmingham, A., Wiemann, S., Beijersbergen, R. L., Hornung, V., & Smith, A. van B. (2015). Advances in CRISPR-Cas9 genome engineering: lessons learned from RNA interference. *Nucleic Acids Research, 43*(7), 3407–3419. http://doi.org/10.1093/nar/gkv226

Barrangou, R. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science, 315*(March), 1709–1712. http://doi.org/10.1126/science.1138140

Bartolomei, M. S., Zemel, S., & Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. *Nature, 351*(6322), 153–5. http://doi.org/10.1038/351153a0

Basu, S., Müller, F., & Sanges, R. (2013). Examples of sequence conservation analyses capture a subset of mouse long non-coding RNAs sharing homology with fish conserved genomic elements. *BMC Bioinformatics, 14 Suppl 7*(Suppl 7), S14. http://doi.org/10.1186/1471-2105-14-S7-S14

Bateman, A., Agrawal, S., Birney, E., Bruford, E. A., Bujnicki, J. M., Cochrane, G., … Zwieb, C. (2011). RNAcentral: A vision for an international database of RNA sequences. *RNA (New York, N.Y.), 17*(11), 1941–6. http://doi.org/10.1261/rna.2750811

Batzer, M. a., Arcot, S. S., Phinney, J. W., Alegria-Hartman, M., Kass, D. H., Milligan, S. M., … Stoneking, M. (1996). Genetic variation of recent Alu insertions in human populations. *Journal of Molecular Evolution, 42*(1), 22–29. http://doi.org/10.1007/BF00163207

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., & Moran, J. V. (2013). LINE-1 Elements in Structural Variation and Disease Christine, *18*(9), 1199–1216. http://doi.org/10.1016/j.micinf.2011.07.011.Innate

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science, 304*(5675), 1321–1325. http://doi.org/10.1126/science.1098119

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., … Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature, 441*(7089), 87–90. http://doi.org/10.1038/nature04696

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B, 57(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bennett, C. F., & Swayze, E. E. (2010). RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annual Review of Pharmacology and Toxicology*, *50*, 259–293. http://doi.org/10.1146/annurev.pharmtox.010909.105654

Bhartiya, D., Pal, K., Ghosh, S., Kapoor, S., Jalali, S., Panwar, B., … Scaria, V. (2013). lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database*, *2013*(0), bat034–bat034. http://doi.org/10.1093/database/bat034

Birney, E. (2006). Ensembl 2006. *Nucleic Acids Research*, *34*(90001), D556–D561. http://doi.org/10.1093/nar/gkj133

Birney, E., Stamatoyannopoulos, J. a, Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. http://doi.org/10.1038/nature05874

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics and Development*, *19*(6), 607–612. http://doi.org/10.1016/j.gde.2009.10.013

Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., … Snyder, M. (2014). Comparative analysis of regulatory information and circuits across distant species. *Nature*, *512*(7515), 453–456. http://doi.org/10.1038/nature13668

Brannan, C. I., Dees, E. C., Ingram, R. S., & Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Molecular and Cellular Biology*, *10*(1), 28–36. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=360709&tool=pmcentrez&rendertype=abstract

Britten, R. J., & Kohne, D. E. (1968). Repeated Sequences in DNA. *SCIENCE*, *161*(4054), 529–540.

Brockdorff, N., Ashworth, a, Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., … Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, *71*(3), 515–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1423610

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*(1), 78–94. http://doi.org/10.1006/jmbi.1997.0951

Buxbaum, A. R., Haimovich, G., & Singer, R. H. (2015). In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol.*, *16*(2), 95–109. http://doi.org/10.1038/nrm3918.In

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–27. http://doi.org/10.1101/gad.17446611

Cargnello, M., Tcherkezian, J., & Roux, P. P. (2015). The expanding role of mTOR in cancer cell growth and proliferation. *Mutagenesis*, *30*(2), 169–176. http://doi.org/10.1093/mutage/geu045

Callinan, P. A., Wang, J., Herke, S. W., Garber, R. K., Liang, P., & Batzer, M. A. (2005). Alu retrotransposition-mediated deletion. *Journal of Molecular Biology*, *348*(4), 791–800. http://doi.org/10.1016/j.jmb.2005.02.043

Carlson, M., Aboyoun, P., & Pages, H. (2015). An Introduction to Genomic Ranges Classes, 1–23. Retrieved from https://bioconductor.org/packages/devel/bioc/vignettes/GenomicRanges/inst/doc/GenomicRangesIntroduction.pdf

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., … Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, *309*(5740), 1559–63. http://doi.org/10.1126/science.1112014

Carninci, P., Nakamura, M., Sato, K., Hayashizaki, Y., & Brownstein, M. J. (2002). Cytoplasmic RNA Extraction from Fresh and Frozen Mammalian Tissues. *BioTechniques*, *33*, 306–309.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., … Hayashizaki, Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discover of new genes. *Genome Res*, *10*, 1431–1432. http://doi.org/10.1101/gr.145100.reflects

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, *38*(6), 626–635. http://doi.org/10.1038/ng1789

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., … Gustincich, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature, 491*(7424), 454–7. http://doi.org/10.1038/nature11508

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science, 34*, 247–278.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. a, Kampa, D., … Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell, 116*(4), 499–509. http://doi.org/10.1016/S0092-8674(04)00127-8

Cenik, C., Cenik, E. S., Byeon, G. W., Grubert, F., Candille, S. I., Spacek, D., … Snyder, M. P. (2015). Integrative analysis of RNA, translation and protein levels reveals distinct regulatory variation across humans. *bioRxiv*, 018572. http://doi.org/10.1101/018572

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., … Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell, 147*(2), 358–69. http://doi.org/10.1016/j.cell.2011.09.028

Chan, W.-L., Huang, H.-D., & Chang, J.-G. (2014). lncRNAMap: A map of putative regulatory functions in the long non-coding transcriptome. *Computational Biology and Chemistry, 50*, 41–49. http://doi.org/10.1016/j.compbiolchem.2014.01.003

Chen, L., Bush, S. J., Tovar-Corona, J. M., Castillo-Morales, A., & Urrutia, A. O. (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Molecular Biology and Evolution, 31*(6), 1402–1413. http://doi.org/10.1093/molbev/msu083

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., & Kim, J. (2014). Analysis of off-target effects of CRISPR Cas-derived RNA-guided endonucleases and nickases sup2, 1–11. http://doi.org/10.1101/gr.162339.113.Freely

Choo, K. H., Tan, T. W., & Ranganathan, S. (2009). A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics, 10 Suppl 1*, S2. http://doi.org/10.1186/1471-2105-10-S15-S2

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., … Mattick, J. S. (2011). The Reality of Pervasive Transcription. *PLoS Biology, 9*(7), e1000625. http://doi.org/10.1371/journal.pbio.1000625

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Hsu, P. D., … Marraffini, L. A. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems, *339*(6121), 819–823. http://doi.org/10.1126/science.1231143.Multiplex

Conley, A. B., Miller, W. J., & Jordan, I. K. (2008). Human cis natural antisense transcripts initiated by transposable elements. *Trends in Genetics*, *24*(2), 53–56. http://doi.org/10.1016/j.tig.2007.11.008

Crampton, N., Bonass, W. a., Kirkham, J., Rivetti, C., & Thomson, N. H. (2006). Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Research*, *34*(19), 5416–5425. http://doi.org/10.1093/nar/gkl668

Cridland, J. M., Thornton, K. R., & Long, A. D. (2015). Gene Expression Variation in Drosophila melanogaster Due to Rare Transposable Element Insertion Alleles of Large Effect. *Genetics*, *199*(1), 85–93. http://doi.org/10.1534/genetics.114.170837

Deininger, P. (2006). Alu elements elements: know the SINEs. *Genomic Disorders: The Genomic Basis of Disease*, 21–34. http://doi.org/10.1007/978-1-59745-039-3_2

Derrien, T., Johnson, R., Bussotti, G., Tanzer, a, Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalogue of human long non-coding RNAs : Analysis of their structure , evolution and expression, 1775–1789. http://doi.org/10.1101/gr.132159.111

Dewannieux, M., Esnault, C., & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, *35*(1), 41–48. http://doi.org/10.1038/ng1223

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, *30*(4), 121–123. http://doi.org/10.1016/j.tig.2014.01.004

Dimitrieva, S., & Bucher, P. (2012). Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics*, *28*(18), 395–401. http://doi.org/10.1093/bioinformatics/bts400

Dorey, F. (2010). The P Value: What is it and what does it tell you? *Clinical Orthopaedics and Related Research*, *468*(8), 2297–2298. http://doi.org/10.1007/s11999-010-1402-9

Dowle, M., Short, T., Lianoglou, S., & Srinivasan, A. (2014). Package "data.table" (Extension of data.frame). Retrieved from https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.pdf

Dinger, M. E., Gascoigne, D. K., & Mattick, J. S. (2011). The evolution of RNAs with multiple functions. *Biochimie*, *93*(11), 2013–8. http://doi.org/10.1016/j.biochi.2011.07.018

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Computational Biology, 4*(11), e1000176. http://doi.org/10.1371/journal.pcbi.1000176

Duret, L., Chureau, C., Samain, S., Weissenbach, J., & Avner, P. (2006). The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. *SCIENCE, 312*(5780), 1653–1655. http://doi.org/10.1177/03063127067078012

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics, 21*(16), 3439–3440. http://doi.org/10.1093/bioinformatics/bti525

Duvernell, D. D., Pryor, S. R., & Adams, S. M. (2004). Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. *Journal of Molecular Evolution, 59*(3), 298–308. http://doi.org/10.1007/s00239-004-2625-8

Ebisuya, M., Yamamoto, T., Nakajima, M., & Nishida, E. (2008). Ripples from neighbouring transcription. *Nature Cell Biology, 10*(9), i. http://doi.org/10.1038/ncb1771

Ekim, B., Magnuson, B., Acosta-Jaquez, H. a., Keller, J. a., Feener, E. P., & Fingar, D. C. (2011). mTOR Kinase Domain Phosphorylation Promotes mTORC1 Signaling, Cell Growth, and Cell Cycle Progression. *Molecular and Cellular Biology, 31*(14), 2787–2801. http://doi.org/10.1128/MCB.05437-11

Eliseeva, I., Vorontsov, I., Babeyev, K., Buyanova, S., Sysoeva, M., Kondrashov, F., & Kulakovskiy, I. (2013). In silico motif analysis suggests an interplay of transcriptional and translational control in mTOR response. *Translation, 1*(2), 18–24. http://doi.org/10.4161/trla.27469

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols, 2*(4), 953–71. http://doi.org/10.1038/nprot.2007.131

Faghihi, M. A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M. P., Nalls, M. a, … Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biology, 11*(5), R56. http://doi.org/10.1186/gb-2010-11-5-r56

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., … Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics, 41*(5), 563–571. http://doi.org/10.1038/ng.368

Feingold, E., Good, P., Guyer, M., Kamholz, S., Liefer, L., Wetterstrand, K., … Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.), 306*, 636–640. http://doi.org/10.1126/science.1105136

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews. Genetics, 9*(5), 397–405. http://doi.org/10.1038/nrg2337

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics, 41*(1), 331–368. http://doi.org/10.1146/annurev.genet.40.110405.090448

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research, 10*(17), 5303–5318. http://doi.org/10.1093/nar/10.17.5303

Fickett, J. W., & Tung, C. S. (1992). Assessment of protein coding measures. *Nucleic Acids Research, 20*(24), 6441–6450. http://doi.org/10.1093/nar/20.24.6441

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research, 42*(D1), D222–D230. http://doi.org/10.1093/nar/gkt1223

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., … Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research, 39*(Database issue), D800–6. http://doi.org/10.1093/nar/gkq1064

Frazier, K. S. (2014). Antisense Oligonucleotide Therapies: The Promise and the Challenges from a Toxicologic Pathologist's Perspective. *Toxicologic Pathology*, 1–12. http://doi.org/10.1177/0192623314551840

Friedman, R., & Hughes, A. L. (2001). Gene duplication and the structure of eukaryotic genomes. *Genome Research, 11*(3), 373–381. http://doi.org/10.1101/gr.155801

Fu, Y., Foden, J. a, Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, pp. 822–6. Nature Publishing Group. http://doi.org/10.1038/nbt.2623

G.D. Schuler, Boguski, M. S., Stewart, E. A., & L.D. Stein. (1996). A gene map of the human genome. *Science, 274*, 540–546.

Gagniuc, P., & Ionescu-Tirgoviste, C. (2012). Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters. *BMC Genomics, 13*(1), 512. http://doi.org/10.1186/1471-2164-13-512

Gal-Mark, N., Schwartz, S., & Ast, G. (2008). Alternative splicing of Alu exons--two arms are better than one. *Nucleic Acids Research, 36*(6), 2012–23. http://doi.org/10.1093/nar/gkn024

Gentleman, R. (2009). R programming for bioinformatics. *Journal of Statistical Software, 29*(Book Review 8). http://doi.org/10.1080/02664760802695884

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80. http://doi.org/10.1186/gb-2004-5-10-r80

Gerstein, M. B., Lu, Z. J., Nostrand, E. L. Van, Cheng, C., Arshinoff, B. I., & Liu, T. (2010). Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science, 330*(December 2010), 1775=1786.

Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., … Waterston, R. (2014). Comparative analysis of the transcriptome across distant species. *Nature, 512*, 445–448. http://doi.org/10.1038/nature13424

Giardine, B., Riemer, C., & Hardison, R. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research, 15*, 1451–1455. http://doi.org/10.1101/gr.4086505.

Giannakakis, A., Zhang, J., Jenjaroenpun, P., Nama, S., Zainolabidin, N., Aau, M. Y., … Guccione, E. (2015). Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. *Scientific Reports, 5*(February), 9737. http://doi.org/10.1038/srep09737

Gilbert, N., Lutz-Prigge, S., & Moran, J. V. (2002). Genomic deletions created upon LINE-1 retrotransposition. *Cell, 110*(3), 315–325. http://doi.org/10.1016/S0092-8674(02)00828-0

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*(8), 850–857. http://doi.org/http://dx.doi.org/10.1016/j.jclinepi.2014.03.012

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology, 11*(8), R86. http://doi.org/10.1186/gb-2010-11-8-r86

Goff, L. a, & Rinn, J. L. (2015). Linking RNA biology to lncRNAs, 1456–1465. http://doi.org/10.1101/gr.191122.115.

Goffeau, a, Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., … Oliver, S. G. (1996). Life with 6000 Genes. *Science, 274*(October), 546–567. http://doi.org/jyu

Goodier, J. L., Mandal, P. K., Zhang, L., & Kazazian, H. H. (2010). Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. *Human Molecular Genetics, 19*(9), 1712–1725. http://doi.org/10.1093/hmg/ddq048

Grabowski, P. J., Seiler, S. R., & Sharp, P. a. (1985). A multicomponent complex is involved in the splicing of messenger RNA precursors. *Cell, 42*(1), 345–53. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3160482

Gray, Y. H. M. (2000). It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. *Trends in Genetics, 16*(10), 461–468. http://doi.org/10.1016/S0168-9525(00)02104-1

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486–489. http://doi.org/10.5812/ijem.3505

Griffiths-Jones, S. (2004). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research, 33*(Database issue), D121–D124. http://doi.org/10.1093/nar/gki081

Griffiths-Jones, S. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research, 34*(90001), D140–D144. http://doi.org/10.1093/nar/gkj112

Grimaldi, G., & Singer, M. F. (1982). A monkey Alu sequence is flanked by 13-base pair direct repeats by an interrupted alpha-satellite DNA sequence. *Proceedings of the National Academy of Sciences of the United States of America, 79*(5), 1497–500. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=346001&tool=pmcentrez&rendertype=abstract

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., … Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature, 458*(7235), 223–7. http://doi.org/10.1038/nature07672

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., … Lander, E. S. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature, 477*(7364), 295–300. http://doi.org/10.1038/nature10398

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., … Regev, A. (2010). Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. *Nature Biotechnology, 28*(5), 503–510. http://doi.org/10.1038/nbt.1633.Ab

Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature, 482*(7385), 339–46. http://doi.org/10.1038/nature10887

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell*, *154*(1), 240–251. http://doi.org/10.1016/j.cell.2013.06.009

Hahn, M. W., & Wray, G. A. (2002). The g-value paradox. *Evolution & Development*, *4*(2), 73–75. http://doi.org/10.1046/j.1525-142X.2002.01069.x

Harmston, N., Baresic, A., & Lenhard, B. (2013). The mystery of extreme non-coding conservation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1632), 20130021. http://doi.org/10.1098/rstb.2013.0021

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., & Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Research*, *30*(5), 1083–1090. http://doi.org/10.1093/nar/30.5.1083

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., … Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, *7 Suppl 1*(Suppl 1), S4.1–9. http://doi.org/10.1186/gb-2006-7-s1-s4

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–74. http://doi.org/10.1101/gr.135350.111

HAVANA - Wellcome Trust Sanger institute. (n.d.). Human and Vertebrate Analysis and Annotation (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/#t_ref). Retrieved October 25, 2015, from http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/#t_ref

He, S., Liu, S., & Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evolutionary Biology*, *11*(1), 102. http://doi.org/10.1186/1471-2148-11-102

Heo, J. B., & Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science (New York, N.Y.)*, *331*(6013), 76–9. http://doi.org/10.1126/science.1197349

Hilder, V. A., Livesey, R. N., Turner, P. C., & Vlad, M. T. (1981). Histone gene number in relation to C-value in amphibians. *Nucleic Acids Research*, *9*(21), 5737–5746.

Hirose, Y., & Manley, J. L. (1998). RNA polymerase II is an essential mRNA poly-adenylation factor. *Nature*, *395*(6697), 93–96. http://doi.org/10.1038/25786

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., … Park, P. J. (2014). Comparative analysis of metazoan chromatin organization. *Nature*, *512*(7515), 449–452. http://doi.org/10.1038/nature13415

Ho, T.-T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., … Mo, Y.-Y. (2015). Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Research*, *43*(3), e17. http://doi.org/10.1093/nar/gku1198

Hodgkin, J. (2001). What does a worm want with 20,000 genes? *Genome Biology*, *2*(11), COMMENT2008. http://doi.org/10.1186/gb-2001-2-11-comment2008

Hogenesch, J. B., Ching, K. a., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., … Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, *106*(4), 413–415. http://doi.org/10.1016/S0092-8674(01)00467-6

Inada, D. C., Bashir, A., Lee, C., Thomas, B. C., Ko, C., Goff, S. a, & Freeling, M. (2003). Conserved noncoding sequences in the grasses. *Genome Research*, *13*(9), 2030–41. http://doi.org/10.1101/gr.1280703

Ingolia, N. T. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, *324*, 218–223. http://doi.org/10.1126/science.1168978.Genome-Wide

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Jack Weiss. (2003). Statistical adjustments based on the false discovery rate (FDR). Retrieved from http://www.unc.edu/courses/2007spring/biol/145/001/docs/lectures/Nov12.html

Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V, Burchard, J., Mao, M., … Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, *21*(6), 635–637. http://doi.org/10.1038/nbt831

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, *33 Suppl*(march), 245–254. http://doi.org/10.1038/ng1089

Jean-Michel Claverie. (2001). What if there are only 30,000 human genes? *Science*, *291*(2), 1255–1257.

Jefferies, H. B. J., Fumagalli, S., Dennis, P. B., Reinhard, C., Pearson, R. B., & Thomas, G. (1997). Rapamycin suppresses 5'TOP mRNA translation through inhibition of p70(s6k). *EMBO Journal*, *16*(12), 3693–3704. http://doi.org/10.1093/emboj/16.12.3693

Jin, J., Liu, J., Wang, H., Wong, L., & Chua, N.-H. (2013). PLncDB: plant long non-coding RNA database. *Bioinformatics*, *29*(8), 1068–1071. http://doi.org/10.1093/bioinformatics/btt107

Johnson, R., & Guigó, R. (2014). The RIDL hypothesis : transposable elements as functional domains of long noncoding RNAs, 959–976. http://doi.org/10.1261/rna.044560.114.

Jurka, J., & Smith, T. (1988). A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(13), 4775–4778. http://doi.org/10.1073/pnas.85.13.4775

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., … Gingeras, T. R. (2004). Novel RNAs identified from a in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, *14*(3), 331–342. http://doi.org/10.1101/gr.2094104

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Svirskas, R., Patel, S., Frise, E., … Celniker, S. E. (2002). The transposable elements of the Drosophila melanogaster euchromatin : a genomics perspective, 1–20.

Kano, H., Godoy, I., Courtney, C., Vetter, M. R., Gerton, G. L., Ostertag, E. M., & Kazazian, H. H. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes and Development*, *23*(11), 1303–1312. http://doi.org/10.1101/gad.1803909

Kapranov, P., Cawley, simon E., Drenkow, J., & Berkiranov, S. (2002). Large-Scale Transcnptional Activity in Chromosomes 21 and 22, *296*(May), 916–920.

Kapranov, P., Cheng, J., Dike, S., Nix, D. a, Duttagupta, R., Willingham, A. T., … Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, *316*(5830), 1484–8. http://doi.org/10.1126/science.1138341

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., … Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*, *9*(4), e1003470. http://doi.org/10.1371/journal.pgen.1003470

Karolchik, D. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(90001), 493D–496. http://doi.org/10.1093/nar/gkh103

Karsch-Mizrachi, I., Nakamura, Y., & Cochrane, G. (2012). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, *40*(D1), D33–D37. http://doi.org/10.1093/nar/gkr1006

Katayama, S., Tomaru, Y., & Kasukawa, T. (2005). Antisense Transcription in the Mammalian Transcriptome. *Science, 309*.

Katti, M. V, Ranjekar, P. K., & Gupta, V. S. (2001). Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. *Mol Biol Evol, 18*(7), 1161–1167.

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature, 332*(6160), 164–6. http://doi.org/10.1038/332164a0

Kelley, D., Hendrickson, D. G., Tenen, D., & Rinn, J. L. (2014). Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biology, 15*(12), 537. http://doi.org/10.1186/s13059-014-0537-5

Kelley, D., & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology, 13*(11), R107. http://doi.org/10.1186/gb-2012-13-11-r107

Kersey, P. J., Staines, D. M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J. C., … Birney, E. (2012). Ensembl Genomes: An integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Research, 40*(D1), 91–97. http://doi.org/10.1093/nar/gkr895

Kim, H. H., Kuwano, Y., Srikantan, S., Lee, E. K., Martindale, J. L., & Gorospe, M. (2009). HuR recruits let-7/RISC to repress c-Myc expression. *Genes & Development, 23*(15), 1743–8. http://doi.org/10.1101/gad.1812509

Kim, V. N. (2006). Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes & Development, 20*(15), 1993–7. http://doi.org/10.1101/gad.1456106

Kim, Y. K., Furic, L., Parisien, M., Major, F., DesGroseillers, L., & Maquat, L. E. (2007). Staufen1 regulates diverse classes of mammalian transcripts. *The EMBO Journal, 26*(11), 2670–81. http://doi.org/10.1038/sj.emboj.7601712

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., & Ger, R. (2003). Antisense Transcripts With FANTOM2 Clone Set and Their Implications for Gene Regulation, 1324–1334. http://doi.org/10.1101/gr.982903.1

Kokocinski, F., Harrow, J., & Hubbard, T. (2010). AnnoTrack--a tracking system for genome annotation. *BMC Genomics, 11*, 538. http://doi.org/10.1186/1471-2164-11-538

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support

vector machine. *Nucleic Acids Research*, *35*(Web Server), W345–W349. http://doi.org/10.1093/nar/gkm391

Kornienko, A. E., Guenzl, P. M., Barlow, D. P., & Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription. *BMC Biology*, *11*(1), 59. http://doi.org/10.1186/1741-7007-11-59

Kramerov, D. a, & Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, *107*(6), 487–95. http://doi.org/10.1038/hdy.2011.43

Krasnov, A., Koskinen, H., Afanasyev, S., & Mölsä, H. (2005). Transcribed Tc1-like transposons in salmonid fish. *BMC Genomics*, *6*, 107. http://doi.org/10.1186/1471-2164-6-107

Lafontaine, D. L., & Tollervey, D. (1998). Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends in Biochemical Sciences*, *23*(10), 383–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9810226

Lamond, A. I., & Spector, D. L. (2003). Nuclear speckles: a model for nuclear organelles. *Nature Reviews. Molecular Cell Biology*, *4*(8), 605–12. http://doi.org/10.1038/nrm1172

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118. http://doi.org/10.1371/journal.pcbi.1003118

Lee, J. T., Davidow, L. S., & Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature Genetics*, *21*(4), 400–4. http://doi.org/10.1038/7734

Lehner, B., Williams, G., Campbell, R. D., & Sanderson, C. M. (2002). Antisense transcripts in the human genome. *Trends Genet*, *18*(2), 63–65. http://doi.org/S0168952502025982 [pii]

Lerat, E., Brunet, F., Bazin, C., & Capy, P. (1999). Is the evolution of transposable elements modular? *Genetica*, *107*(1-3), 15–25. http://doi.org/10.1023/A:1004026821539

Lerat, E., Rizzon, C., & Biemont, C. (2003). Sequence divergence within transposable element families in the Drosophila melanogaster genome. *Genome Res*, *13*(8), 1889–1896. http://doi.org/10.1101/gr.827603.

Li, E., & Zhang, Y. (2014). DNA Methylation in Mammals. *Cold Spring Harb Perspect Biol*, *6*(a019133). http://doi.org/10.1101/cshperspect.a019133

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, S., Cutler, G., Liu, J. J., Hoey, T., Chen, L., Schultz, P. G., … Ling, X. B. (2003). A comparative analysis of HGSC and Celera human genome assemblies and gene sets. *Bioinformatics*, *19*(13), 1597–1605. http://doi.org/10.1093/bioinformatics/btg219

Li, W.-H. (1997). Molecular Evolution. *Sinauer Associates, Sunderland, Massachusetts*.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., & Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics*, *25*(2), 239–40. http://doi.org/10.1038/76126

Lithgow, T. (2000). Targeting of proteins to mitochondria. *FEBS Letters*, *476*(1-2), 22–26. http://doi.org/10.1016/S0014-5793(00)01663-X

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics (Oxford, England)*, *27*(13), i275–82. http://doi.org/10.1093/bioinformatics/btr209

Lin, X., Ruan, X., Anderson, M. G., McDowell, J. a., Kroeger, P. E., Fesik, S. W., & Shen, Y. (2005). siRNA-mediated off-target gene silencing triggered by a 7 nt complementation. *Nucleic Acids Research*, *33*(14), 4527–4535. http://doi.org/10.1093/nar/gki762

Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., … Chen, R. (2005). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*, *33*(Database issue), D112–5. http://doi.org/10.1093/nar/gki041

Liu, G., Zhao, S., Bailey, J. A., Sahinalp, S. C., Alkan, C., Tuzun, E., … Eichler, E. E. (2003). Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Research*, *13*(3), 358–368. http://doi.org/10.1101/gr.923303

Liu, J., Gough, J., & Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genetics*, *2*(4), 529–536. http://doi.org/10.1371/journal.pgen.0020029

Liu, X., Li, D., Zhang, W., Guo, M., & Zhan, Q. (2012). Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. *The EMBO Journal*, *31*(23), 4415–27. http://doi.org/10.1038/emboj.2012.292

Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical Journal*, *417*(1), 15–27. http://doi.org/10.1042/BJ20081501

López-Lastra, M., Rivas, A., & Barría, M. I. (2005). Protein synthesis in eukaryotes: the growing biological relevance of cap-independent translation initiation. *Biological Research*, *38*(2-3), 121–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16238092

Loveland, J. E., Gilbert, J. G. R., Griffiths, E., & Harrow, J. L. (2012). Community gene annotation in practice. *Database, 2012*, 1–8. http://doi.org/10.1093/database/bas009

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*(5), 955–64. http://doi.org/10.1093/nar/25.5.955

Lupski, J. R. (1998). Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics, 14*(10), 417–422. http://doi.org/10.1016/S0168-9525(98)01555-8

Lyle, R., Watanabe, D., te Vruchte, D., Lerchner, W., Smrzka, O. W., Wutz, A., … Barlow, D. P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nature Genetics, 25*(1), 19–21. http://doi.org/10.1038/75546

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics, 154*(1), 459–473. http://doi.org/10.1371/journal.pgen.0040029

Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology, 10*(6), 1–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23696037

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research, 42*(Database issue), D986–92. http://doi.org/10.1093/nar/gkt958

Makarova, J. a, & Kramerov, D. a. (2011). SNOntology: Myriads of novel snoRNAs or just a mirage? *BMC Genomics, 12*(1), 543. http://doi.org/10.1186/1471-2164-12-543

Mao, Y. S., Sunwoo, H., Zhang, B., & Spector, D. L. (2011). Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nature Cell Biology, 13*(1), 95–101. http://doi.org/10.1038/ncb2140

Mariño-Ramírez, L., Lewis, K. C., Landsman, D., & Jordan, I. K. (2005). Transposable elements donate lineage-specific regulatory. *Cytogenet Genome Res, 110 (1-4)*, 333–341. http://doi.org/10.1037/a0030561.Striving

Martens, J. a, Laprade, L., & Winston, F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature, 429*(6991), 571–4. http://doi.org/10.1038/nature02538

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., … Apweiler, R. (2005). PRIDE: The proteomics identifications database. *Proteomics, 5*(13), 3537–3545. http://doi.org/10.1002/pmic.200401303

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., & Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature, 445*(7128), 666–70. http://doi.org/10.1038/nature05519

Matlin, A. J., Clark, F., & Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology, 6*(5), 386–398. http://doi.org/10.1038/nrm1645

Mattick, J. S. (2006). Non-coding RNA. *Human Molecular Genetics, 15*(90001), R17–R29. http://doi.org/10.1093/hmg/ddl046

Mattick, J. S., & Gagen, M. J. (2001). The Evolution of Controlled Multitasked Gene Networks : The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms. *Molecular Biology and Evolution, 18*(9), 1611–1630. http://doi.org/10.1093/oxfordjournals.molbev.a003951

Mcclintock, B. (1951). Chromosome Organization and Genic Expression. *Cold Spring Harb Symp Quant Biol., 16*, 13–17.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., … Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature, 454*(7205), 766–70. http://doi.org/10.1038/nature07107

Messerschmidt, D. M., Knowles, B. B., & Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes and Development, 28*(8), 812–828. http://doi.org/10.1101/gad.234294.113

Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews. Genetics, 10*(3), 155–9. http://doi.org/10.1038/nrg2521

Meyer, I. M. (2008). Predicting novel RNA-RNA interactions. *Current Opinion in Structural Biology, 18*(3), 387–393. http://doi.org/10.1016/j.sbi.2008.03.006

Mirsky, a E., & Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. *The Journal of General Physiology, 34*, 451–462. http://doi.org/10.1085/jgp.34.4.451

Misra, S., Crosby, M. a, Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., … Lewis, S. E. (2002). Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biology, 3*(12), RESEARCH0083. http://doi.org/10.1186/gb-2002-3-12-research0083

Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., … Asai, K. (2009). The Functional RNA Database 3.0: databases to support mining and annotation of functional

RNAs. *Nucleic Acids Research*, *37*(Database), D89–D92. http://doi.org/10.1093/nar/gkn805

Moran, J. V, DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science (New York, N.Y.)*, *283*(5407), 1530–1534. http://doi.org/10.1126/science.283.5407.1530

Mouse Encode Consortium. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol*, *13*(8), 418. http://doi.org/10.1186/gb-2012-13-8-418

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–62. http://doi.org/10.1038/nature01262

Mukhopadhyay, A., Ni, L., & Weiner, H. (2004). A co-translational model to explain the in vivo import of proteins into HeLa cell mitochondria. *The Biochemical Journal*, *382*(Pt 1), 385–392. http://doi.org/10.1042/BJ20040065

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., & C., A. (2008). The Air Noncoding RNA Epigenetically Silences by Targeting Transcription G9a to Chromatin, *322*(5908), 1717–1720.

NCBI Resource Coordinators*. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *43*(D1), D6–D17. http://doi.org/10.1093/nar/gku1130

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., … Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, *505*(7485), 635–640. http://doi.org/10.1038/nature12943

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., … Stamatoyannopoulos, J. a. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics*, *28*(14), 1919–1920. http://doi.org/10.1093/bioinformatics/bts277

Ng, P., Wei, C., Sung, W., Chiu, K. P., Lipovich, L., Ang, C. C., … Ruan, Y. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation, *2*(2), 105–112. http://doi.org/10.1038/NMETH733

Ohno, S. (1970). Evolution by Gene Duplication. Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-86659-3

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symposia in Biology*, *23*, 366–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5065367

Okazaki et al. (2004). Mouse transcriptome: neutral evolution of "non-coding" complementary DNAs (reply). *Nature*, *431*(7010), 1 p following 757; discussion following 757. http://doi.org/10.1038/nature03017

Okazaki Y, Furuno M, Kasukawa T, Adachi J, B. H. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, 563–573.

O'Donnell, K. A., & Boeke, J. D. (2014). Mighty piwis defend the germline against genome intruders. *Cell*, *4*(164), 37–44. http://doi.org/10.1126/scisignal.2001449.Engineering

O'Donnell, K. A., & Burns, K. H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA*, *1*(1), 21. http://doi.org/10.1186/1759-8753-1-21

Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J., & Conklin, D. S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells, 948–958. http://doi.org/10.1101/gad.981002.a

Pang, K. C., Frith, M. C., & Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function, *22*(1), 1–5.

Park, C., Yu, N., Choi, I., Kim, W., & Lee, S. (2014). lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics (Oxford, England)*, *30*(17), 1–6. http://doi.org/10.1093/bioinformatics/btu325

Park, E., & Maquat, L. E. (2013). Staufen-mediated mRNA decay. *Wiley Interdisciplinary Reviews. RNA*, *4*(4), 423–435. http://doi.org/10.1002/wrna.1168.Staufen-mediated

Prak, E. T., & Kazazian Jr, H. H. (2000). Mobile elements and the human genome. *Nature reviews.Genetics*, *1*(2), 134–144.

Patrucco, L., Chiesa, A., Soluri, M. F., Fasolo, F., Takahashi, H., Carninci, P., … Cotella, D. (2015). Engineering mammalian cell factories with SINEUP noncoding RNAs to improve translation of secreted proteins. *Gene*, *569*(2), 287–293. http://doi.org/10.1016/j.gene.2015.05.070

Pauli, A., Rinn, J. L., & Schier, A. F. (2011). Non-coding RNAs as regulators of embryogenesis. *Nature Reviews. Genetics*, *12*(2), 136–49. http://doi.org/10.1038/nrg2904

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., … Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis, 577–591. http://doi.org/10.1101/gr.133009.111.2011

Pereira, V., Enard, D., & Eyre-Walker, A. (2009). The Effect of Transposable Element Insertions on Gene Expression Evolution in Rodents. *PLoS ONE*, *4*(2), e4321. http://doi.org/10.1371/journal.pone.0004321

Polavarapu, N., Mariño-Ramírez, L., Landsman, D., McDonald, J. F., & Jordan, I. K. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics, 9*, 226. http://doi.org/10.1186/1471-2164-9-226

Ponjavic, J., Ponting, C. P., & Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research, 17*(5), 556–65. http://doi.org/10.1101/gr.6036807

Pontier, D. B., & Gribnau, J. (2011). Xist regulation and function explored. *Human Genetics, 130*(2), 223–36. http://doi.org/10.1007/s00439-011-1008-7

Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell, 136*(4), 629–641. http://doi.org/10.1016/j.cell.2009.02.006

Potter, S. C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S. M. J., … Clamp, M. (2004). The Ensembl Analysis Pipeline. *Genome Research, 14*(5), 934–941. http://doi.org/10.1101/gr.1859804

Pruitt, K. D., Harrow, J., Harte, R. a., Wallin, C., Diekhans, M., Maglott, D. R., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research, 19*(7), 1316–1323. http://doi.org/10.1101/gr.080531.108

Qu, Z., & Adelson, D. L. (2012). Evolutionary conservation and functional roles of ncRNA. *Frontiers in Genetics, 3*(October), 205. http://doi.org/10.3389/fgene.2012.00205

Quek, X. C., Thomson, D. W., Maag, J. L. V, Bartonicek, N., Signal, B., Clark, M. B., … Dinger, M. E. (2015). lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research, 43*(Database issue), D168–73. http://doi.org/10.1093/nar/gku988

Quentin, Y. (1994). A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Research, 22*(12), 2222–2227.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England), 26*(6), 841–842. http://doi.org/10.1093/bioinformatics/btq033

Rao, D. D., Vorhies, J. S., Senzer, N., & Nemunaitis, J. (2009). siRNA vs. shRNA: Similarities and differences. Advanced Drug Delivery Reviews, 61(9), 746–759. http://doi.org/10.1016/j.addr.2009.04.004

Reed, R., & Maniatis, T. (1988). The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes & Development, 2*(10), 1268–1276. http://doi.org/10.1101/gad.2.10.1268

Repbase Update - GIRI. (n.d.). Retrieved December 11, 2015, from http://www.girinst.org/repbase/update/search.php?query=B2&querytype=Comments+and+description

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I., & Gage, F. H. (2013). The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *Journal of Neuroscience, 33*(45), 17577–17586. http://doi.org/10.1523/JNEUROSCI.3369-13.2013

Richardson, S. R., Morell, S., & Faulkner, G. J. (2014). L1 Retrotransposons and Somatic Mosaicism in the Brain. *Annual Review of Genetics*, (June), 1–27. http://doi.org/10.1146/annurev-genet-120213-092412

Rinn, J. L. (2014). lncRNAs: linking RNA to chromatin. *Cold Spring Harbor Perspectives in Biology, 6*(8). http://doi.org/10.1101/cshperspect.a018614

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. a, … Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell, 129*(7), 1311–23. http://doi.org/10.1016/j.cell.2007.05.022

Ripley, B., Tierney, L., & Urbanek, S. (2015). Package " parallel " (Support for Parallel Computation), 1–13. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf; https://stat.ethz.ch/R-manual/R-devel/library/parallel/html/parallel-package.html

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., … Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research, 43*(D1), D670–D681. http://doi.org/10.1093/nar/gku1177

Ross Ihaka & Robert Gentleman. (1996). R: A Language for Data Analysis and Graphics.

Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M. L., … Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science (New York, N.Y.), 330*(6012), 1787–97. http://doi.org/10.1126/science.1198374

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., … Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics, 5*(1), 99. http://doi.org/10.1186/1471-2164-5-99

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F., & Stupka, E. (2006). Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology, 7*(7), R56. http://doi.org/10.1186/gb-2006-7-7-r56

Santoro, F., Mayer, D., Klement, R. M., Warczok, K. E., Stukalov, A., Barlow, D. P., & Pauler, F. M. (2013). Imprinted Igf2r silencing depends on continuous Airn lncRNA expression

and is not restricted to a developmental window. *Development (Cambridge, England), 140*(6), 1184–95. http://doi.org/10.1242/dev.088849

Sarma, K., Levasseur, P., Aristarkhov, A., & Lee, J. T. (2010). Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 22196–22201. http://doi.org/10.1073/pnas.1009785107

Saraogi, I., & Shan, S. (2011). Molecular mechanism of co-translational protein targeting by the signal recognition particle. *Traffic (Copenhagen, Denmark), 12*(5), 535–42. http://doi.org/10.1111/j.1600-0854.2011.01171.x

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., … Okada, N. (2008). Possible involvement of SINEs in mammalian-specific brain formation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(11), 4220–4225. http://doi.org/10.1073/pnas.0709398105

Schatz, G., & Dobberstein, B. (1996). Common Principles of Protein Translocation Across Membranes.

Searle, S., Frankish, A., Bignell, A., B., A., T., D., M., D., … T., H. (2010). The GENCODE human gene set. *Genome Biology, 11*(October), 2010. http://doi.org/10.1186/gb-2010-11-s1-p36

Searle, S. M., Gilbert, J., Iyer, V., & Clamp, M. (2004). The otter annotation system. *Genome Res, 14*(5), 963–970. http://doi.org/10.1101/gr.1864804

Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., Reik, W., … Hackett, J. (2013). Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368*(1609), 20110330. http://doi.org/10.1098/rstb.2011.0330

Sela, N., Kim, E., & Ast, G. (2010). The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biology, 11*(6), R59. http://doi.org/10.1186/gb-2010-11-6-r59

Shearwin, K. E., Callen, B. P., & Egan, J. B. (2010). Transcriptional interference – a crash course, *21*(6), 339–345. http://doi.org/10.1016/j.tig.2005.04.009.Transcriptional

Shen, S., Lin, L., Cai, J. J., Jiang, P., Kenkel, E. J., Stroik, M. R., … Xing, Y. (2011). Widespread establishment and regulatory impact of Alu exons in human genes. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 2837–2842. http://doi.org/10.1073/pnas.1012834108

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. http://doi.org/10.1093/nar/29.1.308

She, X., Cheng, Z., Zöllner, S., Church, D. M., & Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nature Genetics*, *40*(7), 909–14. http://doi.org/10.1038/ng.172

Shi, X., Sun, M., Wu, Y., Yao, Y., Liu, H., Wu, G., … Song, Y. (2015). Post-transcriptional regulation of long noncoding RNAs in cancer. *Tumor Biology*, *36*(2), 503–513. http://doi.org/10.1007/s13277-015-3106-y

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., … Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15776–81. http://doi.org/10.1073/pnas.2136655100

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., … Boguski, M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature*, *409*(6822), 922–927. http://doi.org/10.1038/35057141

Silva, J. C., Shabalina, S. a, Harris, D. G., Spouge, J. L., & Kondrashovi, a S. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genetical Research*, *82*(1), 1–18. http://doi.org/10.1017/S0016672303006268

Sleutels, F., Zwart, R., & Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, *415*(6873), 810–813. http://doi.org/10.1038/415810a

Smit, A., & Riggs, A. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Research*, *23*(1), 98–102. http://doi.org/10.1093/nar/23.1.98

Smit, A. F. A., & Green, R. H. & P. (n.d.). RepeatMasker (unpublished). Retrieved October 29, 2015, from http://www.repeatmasker.org/

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., & Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, *484*(7394), 339–44. http://doi.org/10.1038/nature10960

Sorek, R., Ast, G., & Graur, D. (2002). Alu -Containing Exons are Alternatively Spliced. *Genome Research*, *12*, 1060–1067. http://doi.org/10.1101/gr.229302.may

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology, 21*(6), 1973–85. http://doi.org/10.1128/MCB.21.6.1973-1985.2001

Srikanta, D., Sen, S. K., Huang, C. T., Conlin, E. M., Rhodes, R. M., & Batzer, M. A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics, 93*(3), 205–212. http://doi.org/10.1016/j.ygeno.2008.09.016

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology, 14*(2), 103–5. http://doi.org/10.1038/nsmb0207-103

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews. Genetics, 9*(6), 465–76. http://doi.org/10.1038/nrg2341

Swift, H. (1950). The constancy of desoxyribose nucleic acid in plant nuclei. *Proceedings of the National Academy of Sciences of the United States of America, 36*(11), 643–654. http://doi.org/10.1073/pnas.36.11.643

Taft, R. J., & Mattick, J. S. (2004). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology, 5*(1), 25. http://doi.org/10.1186/gb-2003-5-1-p1

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061–1073. http://doi.org/10.1038/nature09534

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature, 408*(6814), 796–815. http://doi.org/10.1038/35048692

The C. elegans Sequencing Consortium. (1998). Genome Sequence Platform of for the Nematode elegans : Investigating Biology. *SCIENCE, 282*(5396), 2012–2018.

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74. http://doi.org/10.1038/nature11247

The International HapMap Consortium. (2003). The International HapMap Project. *Nature, 426*(6968), 789–796. http://doi.org/10.1038/nature02168

The R Development Core Team. (2004). *The R Reference Manual Base Package Volume 2* (Vol. 2). Retrieved from http://www.astropa.inaf.it/CF/DOC/R/vol2.pdf

The UniProt Consortium. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Research, 35*(Database), D193–D197. http://doi.org/10.1093/nar/gkl929

Thomas, C. a. (1971). The genetic organization of chromosomes. *Annual Review of Genetics, 5*, 237–256. http://doi.org/10.1146/annurev.ge.05.120171.001321

Thoreen, C. C., Chantranupong, L., Keys, H. R., Wang, T., Gray, N. S., & Sabatini, D. M. (2012). A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*, *486*(7396), 109–113. http://doi.org/10.1038/nature11083

Thornburg, B. G., Gotea, V., & Makałowski, W. (2006). Transposable elements as a significant source of transcription regulating signals. *Gene*, *365*(1-2 SPEC. ISS.), 104–110. http://doi.org/10.1016/j.gene.2005.09.036

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. http://doi.org/10.1038/nbt.1621

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., … Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*, *39*(6), 925–38. http://doi.org/10.1016/j.molcel.2010.08.011

Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M., & Nakagawa, S. (2011). Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes to Cells : Devoted to Molecular & Cellular Mechanisms*, *16*(5), 479–90. http://doi.org/10.1111/j.1365-2443.2011.01502.x

Tuck, A. C., & Tollervey, D. (2013). A Transcriptome-wide Atlas of RNP Composition Reveals Diverse Classes of mRNAs and lncRNAs. *Cell*, *154*(5), 996–1009. http://doi.org/10.1016/j.cell.2013.07.047

Turlan, C., Loot, C., & Chandler, M. (2004). IS 911 partial transposition products and their processing by the Escherichia coli RecG helicase, *53*, 1021–1033. http://doi.org/10.1111/j.1365-2958.2004.04165.x

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2012). Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell*, *151*(3), 684–686. http://doi.org/10.1016/j.cell.2012.10.002

van Bakel, H., Nislow, C., Blencowe, B. J., & Hughes, T. R. (2010). Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLoS Biology*, *8*(5), e1000371. http://doi.org/10.1371/journal.pbio.1000371

Van De Lagemaat, L. N., Gagnier, L., Medstrand, P., & Mager, D. L. (2005). Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Research*, *15*(9), 1243–1249. http://doi.org/10.1101/gr.3910705

van Wolfswinkel, J. C., & Ketting, R. F. (2010). The role of small non-coding RNAs in genome stability and chromatin organization. *Journal of Cell Science, 123*(Pt 11), 1825–39. http://doi.org/10.1242/jcs.061713

Varshney, D., Vavrova-Anderson, J., Oler, A. J., Cowling, V. H., Cairns, B. R., & White, R. J. (2015). SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. *Nature Communications, 6*, 6569. http://doi.org/10.1038/ncomms7569

Veniaminova, N. A., Vassetzky, N. S., & Kramerov, D. A. (2007). B1 SINEs in different rodent families. *Genomics, 89*(6), 678–686. http://doi.org/10.1016/j.ygeno.2007.02.007

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–51. http://doi.org/10.1126/science.1058040

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., … Mestdagh, P. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research, 41*(D1), D246–D251. http://doi.org/10.1093/nar/gks915

Walter, P., & Blobel, G. (1982). Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature, 299*, 691 – 698. http://doi.org/10.1017/CBO9781107415324.004

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., … Wong, G. K.-S. (2004). Mouse transcriptome: Neutral evolution of "non-coding" complementary DNAs. *Nature, 431*(7010), 14–15. http://doi.org/10.1038/nature03016

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., … Chang, H. Y. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature, 472*(7341), 120–4. http://doi.org/10.1038/nature09819

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research, 41*(6), e74–e74. http://doi.org/10.1093/nar/gkt006

Washietl, S., Findeiss, S., Müller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., … Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA (New York, N.Y.), 17*(4), 578–94. http://doi.org/10.1261/rna.2536111

Watts, J., & Corey, D. (2012). Gene silencing by siRNAs and antisense oligonucleotides in the laboratory and the clinic. *The Journal of Pathology, 226*(2), 365–379. http://doi.org/10.1002/path.2993.Gene

Wheeler, D. L. (2004). Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Research*, *32*(90001), 35D–40. http://doi.org/10.1093/nar/gkh073

Wickham, H. (2009). *Polishing your plots for publication - ggplot2: Elegant Graphics for Data Analysis (Use R!)*. New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-98141-3

Wickham, H. (2014). Package "reshape2" (Flexibly Reshape Data: A Reboot of the Reshape Package). Retrieved from https://cran.r-project.org/web/packages/reshape2/reshape2.pdf

Williams, C. C., Jan, C. H., & Weissman, J. S. (2014). Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science*, *346*(6210), 748–751. http://doi.org/10.1126/science.1257522

Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world, 1494–1504. http://doi.org/10.1101/gad.1800909.complexity.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., … Elgar, G. (2005). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology*, *3*(1), e7. http://doi.org/10.1371/journal.pbio.0030007

Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews. Genetics*, *12*(8), 542–53. http://doi.org/10.1038/nrg3035

Wutz, A., Rasmussen, T. P., & Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics*, *30*(2), 167–74. http://doi.org/10.1038/ng820

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., … Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Research*, *42*(D1), D98–D103. http://doi.org/10.1093/nar/gkt1222

Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Curtis, J., … Wallender, E. K. (2011). Empiricat Analysis of Transcriptionat Activity in the. *Advancement Of Science*.

Yamashita, R., Suzuki, Y., Takeuchi, N., Wakaguri, H., Ueda, T., Sugano, S., & Nakai, K. (2008). Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics. *Nucleic Acids Research*, *36*(11), 3707–3715. http://doi.org/10.1093/nar/gkn248

Yanagiya, A., Suyama, E., Adachi, H., Svitkin, Y. V., Aza-Blanc, P., Imataka, H., … Sonenberg, N. (2012). Translational Homeostasis via the mRNA Cap-Binding Protein, eIF4E. *Molecular Cell*, *46*(6), 847–858. http://doi.org/10.1016/j.molcel.2012.04.004

Yao, Y., Jin, S., Long, H., Yu, Y., Zhang, Z., Cheng, G., … Wu, Q. (2015). RNAe: an effective method for targeted protein translation enhancement by artificial non-coding RNA with SINEB2 repeat. *Nucleic Acids Research*, 1–18. http://doi.org/10.1093/nar/gkv125

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., … Flicek, P. (2015). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, *31*(1), 143–145. http://doi.org/10.1093/bioinformatics/btu613

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., … Rotman, G. (2003). Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, *21*(4), 379–386. http://doi.org/10.1038/nbt808

Yoon, J. H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., … Gorospe, M. (2012). LincRNA-p21 Suppresses Target mRNA Translation. *Molecular Cell*, *47*(4), 648–655. http://doi.org/10.1016/j.molcel.2012.06.027

Yoon, J.-H., Abdelmohsen, K., & Gorospe, M. (2013). Posttranscriptional Gene Regulation by Long Noncoding RNA. *Journal of Molecular Biology*, *425*(19), 3723–3730. http://doi.org/10.1016/j.jmb.2012.11.024

Yotova, I. Y., Vlatkovic, I. M., Pauler, F. M., Warczok, K. E., Ambros, P. F., Oshimura, M., … Barlow, D. P. (2008). Identification of the human homolog of the imprinted mouse Air non-coding RNA. *Genomics*, *92*(6), 464–73. http://doi.org/10.1016/j.ygeno.2008.08.004

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., … Mouse, E. C. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355–364. http://doi.org/10.1038/nature13992

Zepeda, A., Arias, C., Flores-Jasso, F., & Vaca, L. (2013). Chapter 17 – RNA Imaging: Tracking in Real-Time RNA Transport in Neurons Using Molecular Beacons and Confocal Microscopy. In *Methods in Cell Biology* (Vol. 113, pp. 361–389). http://doi.org/10.1016/B978-0-12-407239-8.00017-3

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The Ensembl Regulatory Build. *Genome Biology*, *16*(1), 56. http://doi.org/10.1186/s13059-015-0621-5

Zhang, Y., Wang, X., & Kang, L. (2011). A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics*, *27*(6), 771–776. http://doi.org/10.1093/bioinformatics/btr016

Zhang, J., Zuo, T., Wang, D., & Peterson, T. (2014). Transposition-mediated DNA re-replication in maize. *eLife*, *3*, e03724. http://doi.org/10.7554/eLife.03724

Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J., & Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (New York, N.Y.)*, *322*(5902), 750–6. http://doi.org/10.1126/science.1163045

Zong, X., Huang, L., Tripathi, V., Peralta, R., Freier, S. M., Guo, S., & Prasanth, K. V. (2015). Knockdown of nuclear-retained long noncoding RNAs using modified DNA antisense oligonucleotides. *Methods in Molecular Biology (Clifton, N.J.)*, *1262*, 321–31. http://doi.org/10.1007/978-1-4939-2253-6_20

Zucchelli, S., Cotella, D., Takahashi, H., Carrieri, C., Cimatti, L., Fasolo, F., … Gustincich, S. (2015b). SINEUPs: A new class of natural and synthetic antisense long non-coding RNAs that activate translation. *RNA Biol*, *12*(8), 771–779. http://doi.org/10.1080/15476286.2015.1060395

Zucchelli, S., Fasolo, F., Russo, R., Cimatti, L., Patrucco, L., Takahashi, H., … Gustincich, S. (2015a). SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. *Frontiers in Cellular Neuroscience*, *9*(May), 1–12. http://doi.org/10.3389/fncel.2015.00174

Zwart, R., Sleutels, F., Wutz, A., Schinkel, A. H., & Barlow, D. P. (2001). Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. *Genes & Development*, *15*(18), 2361–2366. http://doi.org/10.1101/gad.206201.A